DILS 2014

10th International Conference on Data Integration in the Life Sciences

Poster and Demo Papers

Editors: Helena Galhardas and Erhard Rahm

July 2014

Preface

This volume contains the poster and demo papers presented at DILS 2014: 10th International Conference on Data Integration in the Life Sciences held on July 16-17, 2014 in Lisbon. DILS is an interdisciplinary conference series focussing on data integration and data analysis in life science disciplines including bioinformatics, biomedicine, clinical information systems and biodiversity. The conference series started in 2004 in Leipzig and so far took place in five European countries (Germany, UK, France, Sweden, Portugal), in the USA (three times) and in Canada.

This volume contains the poster and demo papers of DILS 2014. The submissions were limited to 4 pages and are meant to cover ongoing work and prototypes to demonstrate. Each submission was reviewed by 2 program committee members. The committee decided to accept 13 poster and demo papers which are included in this volume. The accepted papers cover different topics such as biomedical ontologies, linked data and biomedical integration platforms and tools. In addition to the poster and demo track DILS 2014 featured two keynotes and 14 research and application papers that are published in the conference proceedings (Springer Lecture Notes in Bioinformatics, No. 8574).

As the event co-chairs and editors of this volume, we would like to thank all authors who submitted papers, as well as the Program Committee members who reviewed the poster and demo submissions in addition to the submissions for the main program. Special thanks go to INESC-ID and Instituto Superior Técnico, University of Lisbon for providing us with the facilities to organize and run the event. We also would like to thank FCT (*Fundação para a Ciência e Tecnologia*) for the financial support provided, in particular through the excellence research network "DataStorm - Large-Scale Data Management in Cloud Environments". Finally, our thanks go to the local Organizing Committee, Ana Teresa Freitas, José Borbinha, José Leal, Mário J. Silva and Pedro T. Monteiro, our Webmaster, João L. M. Pereira, and our administrative staff from INESC-ID, Manuela Sado and Sandra Sá.

July 2014

Helena Galhardas
Erhard Rahm

Program Committee Chair

Helena Galhardas	INESC-ID and Instituto Superior Técnico, University of Lisbon, Portugal
Erhard Rahm	University of Leipzig, Germany

Program Committee

Christopher Baker	University of New Brunswick, Canada
Kenneth J Barker	IBM, USA
Olivier Bodenreider	NIH, USA
João Carriço	IMM, Portugal
Claudine Chaouiya	IGC, Portugal
James Cimino	National Library of Medicine, USA
Luis Pedro Coelho	EMBL, Germany
Sarah Cohen-Boulakia	LRI, University of Paris-Sud 11, France
Francisco Couto	Faculty of Sciences, University of Lisbon, Portugal
Alexandre Francisco	INESC-ID and Instituto Superior Técnico, University of Lisbon, Portugal
Juliana Freire	NYU-Poly, USA
Christine Froidevaux	LRI University of Paris-Sud 11, France
Hasan Jamil	University of Idaho, USA
Graham Kemp	Chalmers University of Technology, Sweden
Toralf Kirsten	University of Leipzig, Germany
Birgitta König-Ries	Institut für Informatik Friedrich-Schiller-Universität Jena, Germany
Patrick Lambrix	Linköping University, Sweden
Adam Lee	University of Maryland and National Library of Medicine, USA
Mong Li Lee	National University of Singapore, Singapore
Ulf Leser	Humboldt-Universität zu Berlin, Germany
Bertram Ludaescher	University of California, USA
Paolo Missier	Newcastle University, UK
Norman Paton	University of Manchester, UK
Cédric Prusky	CRP Henri Tudor, Luxemburg
Uwe Scholz	IPK Gatersleben, Germany
Maria Esther Vidal	Universidad Simón Bolívar, Venezuela
Dagmar Waltemath	University of Rostock, Germany

Additional Reviewers

João D. Ferreira	Faculty of Sciences, University of Lisbon, Portugal

Anika Groß	University of Leipzig, Germany

DILS Steering Committee

Christopher Baker	University of New Brunswick, Canada
Sarah Cohen-Boulakia	LRI, University of Paris-Sud 11, France
Graham Kemp	Chalmers University of Technology, Sweden
Ulf Leser	Humboldt-Universität zu Berlin, Germany
Paolo Missier	Newcastle University, UK
Norman Paton	University of Manchester, UK
Erhard Rahm	University of Leipzig, Germany
Louiqa Raschid	University of Maryland, USA

Organizing Committee

José Borbinha	INESC-ID and Instituto Superior Técnico, University of Lisbon, Portugal
Ana Teresa Freitas	INESC-ID and Instituto Superior Técnico, University of Lisbon, Portugal
José Leal	Instituto Gulbenkian de Ciência, Portugal
Pedro T. Monteiro	INESC-ID, Portugal
Mário J. Silva	INESC-ID and Instituto Superior Técnico, University of Lisbon, Portugal

Webmaster

João L. M. Pereira	INESC-ID and Instituto Superior Técnico, University of Lisbon, Portugal

Table of Contents

Bioinformatics for Operational Precision Medicine in Oncology: application to the SHIVA clinical trial 1
Philippe Hupé

The GigaSolution to data publication, reuse and integration 4
Christopher Hunter, Peter Li, Xiao Si Zhe, Robert Davidson, Laurie Goodman and Scott C. Edmunds

The MicroScope platform: from data integration to a rule-based system for massive and high-quality microbial genome annotation 7
Jonathan Mercier, Alexandre Renaux, Adrien Josso, Aurélie Genin-Lajus, E'Krame Jacoby, François Le Fevre, Guillaume Albini, Claude Scarpelli, Claudine Medigue and David Vallenet

The e!DAL Java API: Sharing and Citing Research Data in Life Sciences 10
Daniel Arend, Jinbo Chen, Christian Colmsee, Steffen Flemming, Denny Hecht, Uwe Scholz and Matthias Lange

Identifiers.org: integration tool for heterogeneous datasets 14
Camille Laibe, Sarala Wimalaratne, Nick Juty, Nicolas Le Novère and Henning Hermjakob

Search QTL candidate genes by LAILAPS information retrieval system . . 16
Jinbo Chen, Daniel Arend, Christian Colmsee, Maria Esch, Uwe Scholz and Matthias Lange

PRIDE Proteomes: a Condensed View of the Plethora of Public Proteomics Data Available in the PRIDE Repository 21
Noemi Del Toro Ayllón, Florian Reisinger, Joseph M. Foster, Javier Contell, Antonio Fabregat Mundo, Pau Ruiz Safont, Henning Hermjakob and Juan A Vizcaíno

A Machine Learning based Natural Language Interface for a database of medicines 25
Ricardo Ferrão, Helena Galhardas and Luísa Coheur

AgreementMakerLight: A Scalable Automated Ontology Matching System 29
Daniel Faria, Catia Pesquita, Emanuel Santos, Isabel F. Cruz and Francisco Couto

Towards a Linked Biology - An integrated perspective of phenotypes and phylogenetic trees 33
Eduardo Miranda, Anaïs Grand, Régine Vignes Lebbe and André Santanchè

Developments and perspectives for CABRI Web Services 37
Daniele Pierpaolo Colobraro and Paolo Romano

Towards a Plant Experimental Assay Ontology 41
Nuno D. Mendes, Pedro T. Monteiro, Cátia Vaz and Inês Chaves

Identifying, interpreting, and communicating changes in XML-encoded models of biological systems 45
Martin Scharm, Olaf Wolkenhauer and Dagmar Waltemath

Bioinformatics for Precision Medicine in Oncology: principles and application to the SHIVA clinical trial

Nicolas Servant[1,2,3]*, Julien Roméjon[1,2,3], Pierre Gestraud[1,2,3], Philippe La Rosa[1,2,3], Georges Lucotte[1,2,3], Séverine Lair[1,2,3], Virginie Bernard[4], Bruno Zeitouni[1,2,3], Fanny Coffin[1,2,3], Gérôme Jules-Clément[1,2,3,10], Florent Yvon[1,2,3], Alban Lermine[1,2,3], Patrick Poullet[1,2,3], Stéphane Liva[1,2,3], Stuart Pook[1,2,3], Tatiana Popova[1,9], Camille Barette[1,2,3,5], François Prud'homme[1,2,3,4,5], Jean-Gabriel Dick[5], Maud Kamal[6], Christophe Le Tourneau[2,3,7], Emmanuel Barillot[1,2,3] and Philippe Hupé[1,2,3,8]*

1 Institut Curie, Paris, France
2 INSERM U900, Paris, France
3 Mines ParisTech, Fontainebleau, France
4 Institut Curie, Sequencing facility ICGex, Paris, France
5 Institut Curie, Informatic Department, Paris, France
6 Institut Curie, Translational Research Department, Paris, France
7 Institut Curie, Department of Medical Oncology, Paris, France
8 CNRS UMR144, Paris, France, 9 INSERM U830, Paris, France
9 INSERM U830, Paris, France
10 INSERM U932, Paris, France

Correspondence: Philippe Hupé, Plateforme de Bioinformatique, Unité 900 : Institut Curie - INSERM - Mines ParisTech, UMR144: Institut Curie - CNRS, 26, rue d'Ulm, 75248 Paris Cedex 05 – FRANCE. philippe.hupe@curie.fr

Introduction

Personalized medicine requires a strong interdisciplinary collaboration between several stakeholders covering a large range of expertise from medical, clinical, biological, translational, biotechnological and bioinformatics fields. The variety of actors and cultures, together with the diversity of constraints make it very challenging to apply personalized medicine in daily clinical practice, to meet expected short deadlines to deliver the results. Personalized medicine strongly depends on our ability to collect, disseminate and process complex information. Indeed, every stakeholder produces information during the healthcare pathway at different time and different places and all these informations have to be gathered, integrated and summarized in a digested report to make easy the therapeutic decision according to rules defined a priori.

Materials and Methods

The KDI system

To tackle the challenge of data integration, we have developed a dedicated information system named KDI (Knowledge and Data Integration) able to handle the heterogeneity and the complexity of the data (Servant et al., 2014). The KDI system ensures information sharing, cross-software interoperability, automatic data extraction, and secure data transfer. Using state-of-the-art informatics technologies, KDI acts as a hub to allow all the data being referenced such that it knows exhaustively which data is available for a given patient and where the raw and processed data are physically stored.

The KDI system consists of different modules dedicated to the storage, processing, analysis and visualization of each type of data (clinical, biological, microarray, NGS, etc.). High modularity associated with an efficient interoperability makes our system able to retrieve any relevant information. To facilitate the developments of these modules, we have retained a classical n-tiers architecture implemented with the JAVA/J2EE language. The core of each module of the KDI system can be presented as the association of different layers (Figure 3B).

Data layer. Data are stored in a relational database using the Entity-Attribute-Value (EAV) pattern. This conceptual modeling provides a data model plasticity required to handle the heterogeneity and the scalability of the variables of interest. Therefore, with EAV modeling, same concepts managed by different projects (with specific requirements by project) can be stored in a unique database without any modification of the data model. MySQL has been chosen as database provider for all web applications of the system. Complementary solutions such as NoSQL databases are currently evaluated for particular requirements (ontologies storage, specific queries, etc.).

Data access layer. Data access is supported by the DAO (Data Access Object) pattern. By using HibernateDaoSupport superclass provided by Spring Framework, we promote the standardization of database access for all standard queries (findAll, findById, save, delete). Moreover, Hibernate mapping through JPA annotations associated with use of Hibernate Criteria provides a homogeneous frame for this critical layer. Database sessions and transactional aspects are also delegated to Spring Framework.

Business layer. Business core of our web applications has two main objectives: i) provide structured data for presentation layer, and ii) make data available for remote and secured access by other applications and technical users. Standard services are developed using core functionalities of Spring framework (Aspect-Oriented Programming - AOP, Inversion of Control - IoC, JavaBeans Factory). Web services are published (server side) and invoked (client side) through Apache CXF framework. To respect Web Services Security (WS-Security) standards, we use the Apache WSS4J project provided by CXF (with interceptors chain process) to set up a username token authentication on each web application in the system.

Front-end layer. Presentation layer is based on JSF (Java Server Faces) which is a component oriented framework for building user interfaces for web applications. To enrich the basic component set provided by JSF, we use additional component libraries such as Apache Trinidad and Primefaces. By this systematic approach for each user interface, we aim to build a visual identity, ergonomic, easily usable, for the whole information system. All data available within KDI can be browsed and retrieved from a user-friendly bioinformatics web portal.

Client layer. This layer represents the web browser through which end-users access KDI system.

The SHIVA clinical trial design

The SHIVA clinical trial is a randomized proof-of-concept phase II trial comparing molecularly targeted therapy based on tumour molecular profiling versus conventional therapy in patients with refractory cancer (Le Tourneau et al., 2012, 2014). For each patient, a biopsy from the metastasis is performed and the molecular profiles are assessed using both the Cytoscan HD technology (Affymetrix) for the detection of DNA copy number alterations and loss of heterozygosity (LOH), and the Ion Torrent™ PGM sequencing technology (Life Technology) for the detection of somatic mutations. Immunohistochemistry (IHC) is used for the assessment of hormone receptor status, including oestrogen, progesterone and androgen receptors, as well as for the validation of focal gene amplifications detected with Cytoscan HD. DNA copy number amplifications (Affymetrix Cytoscan HD microarray) and mutations (next-generation sequencing with IonTorrent) in a subset of 76 genes along with biomarkers detected by IHC are considered for the decision-making.

Data preprocessing

DNA copy number amplifications (Affymetrix Cytoscan HD microarray) and mutations (next-generation sequencing with IonTorrent) are analyzed by in-house *ad-hoc* pipelines. The raw data as long as the results of the pipelines are stored in the KDI system. Clinical data are sored in the system too.

Integrative analysis: the report for the Molecular Biology Board (MBB)

The last step of the bioinformatics workflow is the production a technical report for the MBB. This task is crucial and must be complete and precise on one hand, and summarized on the other to allow a quick decision of the board. To answer this need, a report is automatically generated for each patient from the data stored in KDI. This report first presents the clinical information of the patient and the overall molecular profiles per gene, with the DNA copy number alterations, LOH (Loss of Heterozigosity) status, and number of mutations. This first section provides the MBB with a rapid overview of all detected alterations. If needed, the MBB can also have access to more detailed results, with graphical views of the copy number profiles for each gene, as well as the list of mutations with detailed annotation as previously described. This name-blinded technical report is sent to the members of the MBB for scientific validation and prioritization of the identified molecular abnormalities.

Results

We have developed a seamless information system named KDI that fully supports the essential bioinformatics requirements for PM. The system allows management and analysis of clinical information, classical biological data as well as high-throughput molecular profiles. It can deliver in real-time information to be used by the medical and biological staff for therapeutic decision-making. KDI makes it possible to share information and communicate reports and results across numerous stakeholders, representing a large continuum of expertise from medical, clinical, biological, translational, technical and biotechnological know-hows. The system relies on state-of-the-art informatic technologies allowing cross-software interoperability, automatic data extraction, quality control and secure data transfer. KDI has been successfully used in the framework of the SHIVA clinical trial for more than 18 months. As of June 2014, 730 patients have been included and 152 randomized in the SHIVA trial. KDI is used for other trials in the framawork of european FP7 projects (RAIDs for cervival cancer – http://www.raids-fp7.eu/ ; MAARS for allergy and autoimmune diseases related to skin – http://www.maars.eu/).The KDI is curently used to manage all the high-throughput data in Institut Curie.

References

Le Tourneau C, Kamal M, Trédan O, Delord JP, Campone M, Goncalves A, Isambert N, Conroy T, Gentien D, Vincent-Salomon A, Pouliquen AL, Servant N, Stern MH, Le Corroller AG, Armanet S, Rio Frio T, Paoletti X. Designs and challenges for personalized medicine studies in oncology: focus on the SHIVA trial. Target Oncol. 2012 Dec;7(4):253-65.

Servant N, Roméjon J, Gestraud P, La Rosa P, Lucotte G, Lair S, Bernard V, Zeitouni B, Coffin F, Jules-Clément G, Yvon F, Lermine A, Poullet P, Liva S, Pook S, Popova T, Barette C, Prud'homme F, Dick J, Kamal M, Le Tourneau C, Barillot E and Hupé P. Bioinformatics for Precision Medicine in Oncology: principles and application to the SHIVA clinical trial. 2014. Front. Genet. 5:152. doi: 10.3389/fgene.2014.00152

Le Tourneau C, Paoletti X, Servant N, Bièche I, Gentien D, Rio Frio T, Vincent-Salomon A, Servois V, Romejon J, Mariani O,1 Bernard V, Hupé P, Pierron G, Mulot M, Callens C, Wong J, Mauborgne C, Rouleau E, Reyes C, Leroy Q, Henri E, Gestraud P, La Rosa P, Escalup L, Mitry E, Trédan O, Delord JP, Campone M, Goncalves A, Isambert N, Gavoille C , Kamal M. Randomized proof-of-concept phase II trial comparing therapy based on tumor molecular profiling versus conventional therapy in patients with refractory cancer: Results of the feasibility part of the SHIVA trial. Br J Cancer. 2014 Apr 24. doi: 10.1038/bjc.2014.211.

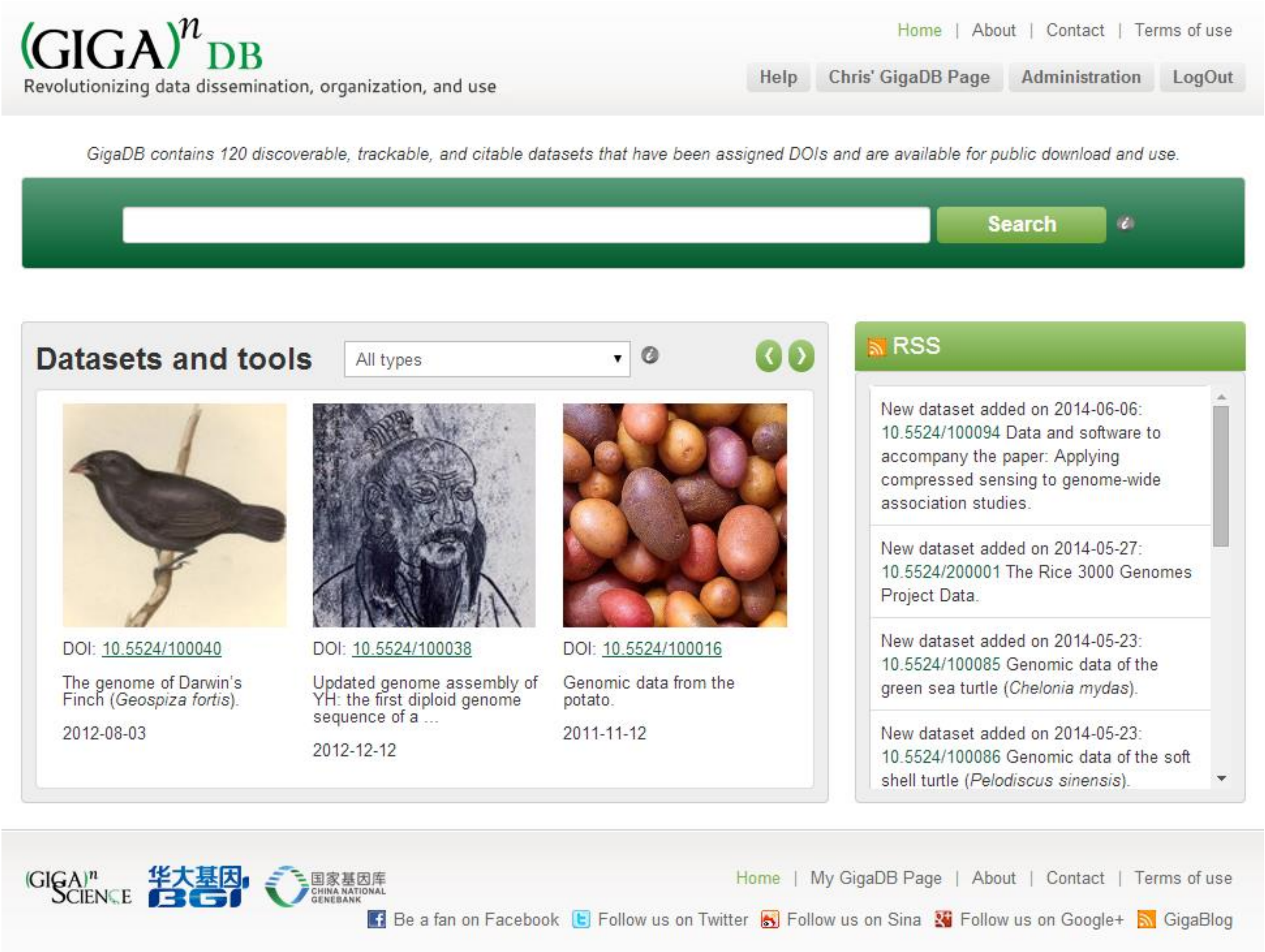
The GigaSolution to data publication, reuse and integration.

Christopher I Hunter, Peter Li, Xiao Si Zhe, Robert Davidson, Laurie Goodman & Scott C Edmunds

Affiliation: GigaScience, BGI-HK Research Institute, 16 Dai Fu Street, Tai Po Industrial Estate, Hong Kong SAR, China.

Correspondence to chris@gigasciencejournal.com

To meet the needs of a new generation of biological and biomedical research in the era of “big-data, BGI and BioMed Central have formed a unique partnership to publish the journal *GigaScience. GigaScience* is a novel publishing platform that combines the open-access article publishing expertise of BMC with the bioinformatics expertise and extensive computational storage space at BGI. The journal’s affiliated database, *Giga*DB (Figure 1), serves as a repository that hosts the data and tools associated with *GigaScience* publications. It also provides a rapid data release mechanism for datasets that are not associated with *GigaScience* articles that have not previously been published elsewhere by giving each the dataset a DOI, making them citable in a standard (and countable) manner in the reference section of papers that use these data.



Revolutionizing data dissemination, organization, and use

GigaDB contains 120 discoverable, trackable, and citable datasets that have been assigned DOIs and are available for public download and use.

DOI: 10.5524/100040

The genome of Darwin's Finch (*Geospiza fortis*).

2012-08-03

DOI: 10.5524/100038

Updated genome assembly of YH: the first diploid genome sequence of a ...

2012-12-12

DOI: 10.5524/100016

Genomic data from the potato.

2011-11-12

New dataset added on 2014-06-06: 10.5524/100094 Data and software to accompany the paper: Applying compressed sensing to genome-wide association studies.

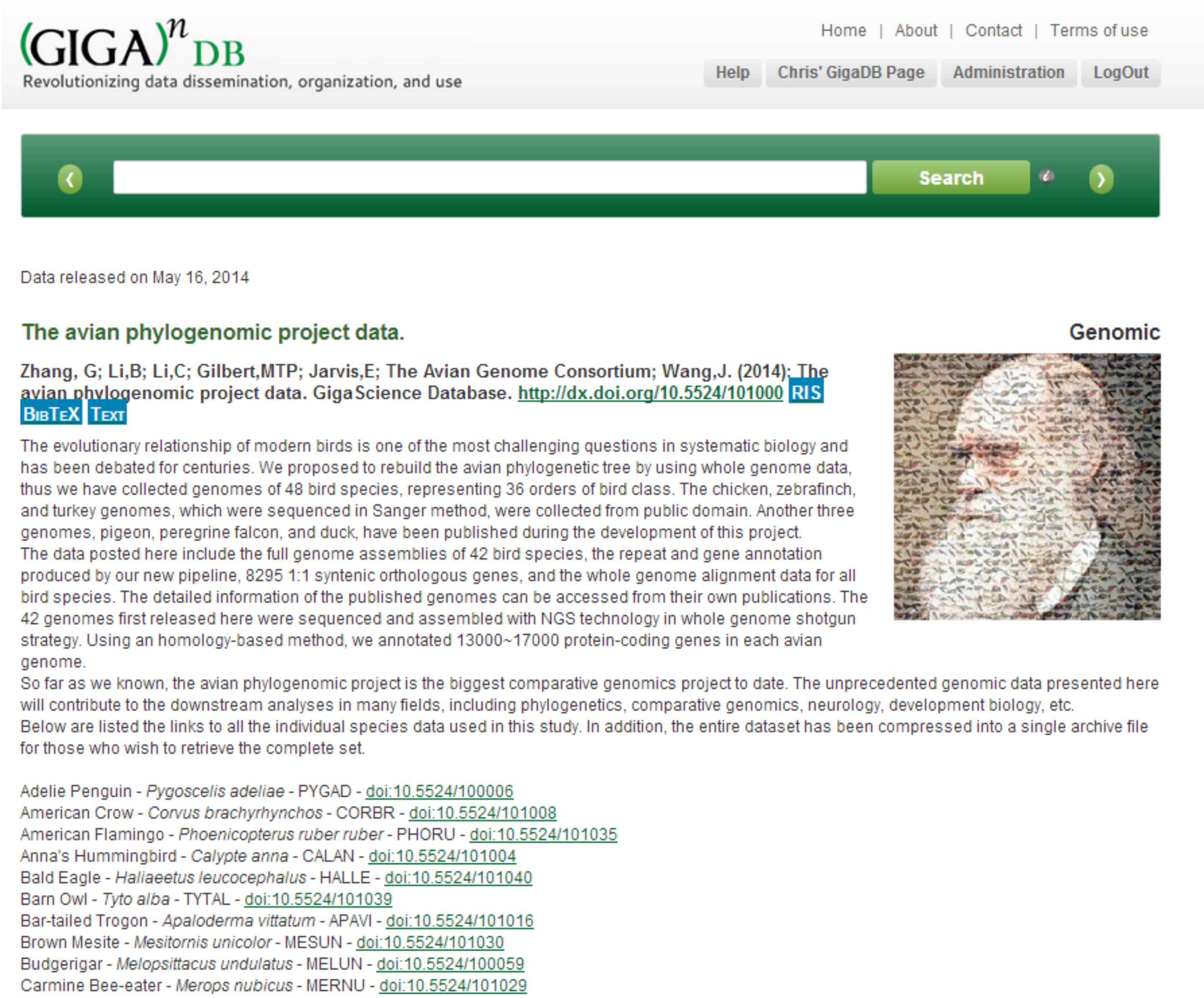
New dataset added on 2014-05-27: 10.5524/200001 The Rice 3000 Genomes Project Data.

New dataset added on 2014-05-23: 10.5524/100085 Genomic data of the green sea turtle (*Chelonia mydas*).

New dataset added on 2014-05-23: 10.5524/100086 Genomic data of the soft shell turtle (*Pelodiscus sinensis*).

Figure 1. The home page of GigaDB website

In its first 18 months, over 100 datasets (>20 TB in size) have been made available in *Giga*DB — all under a CC0 Waiver (the most open sharing waiver available). These datasets include the first ~50 bird genomes from an avian phylogenomics study (Figure 2)(Zhang *et al.*, 2014) some of these datasets were made available before they were published in scientific journals. *Giga*DB will also host data from the Rice 3000 genomes project (The 3000 rice genomes project *et al*,. 2014), the 10,000 genome project (G10K), 1000 plant transcriptomes project, as well other smaller scale genome projects, and numerous non-genomic datasets (e.g. imaging, proteomic, metabolomics, etc.), some of which currently have no formal community data repository.


(GIGA)ⁿDB
Revolutionizing data dissemination, organization, and use
Home | About | Contact | Terms of use
Help
Chris' GigaDB Page
Administration
LogOut
Search
Genomic

Data released on May 16, 2014

The avian phylogenomic project data.

Zhang, G; Li,B; Li,C; Gilbert,MTP; Jarvis,E; The Avian Genome Consortium; Wang,J. (2014): The avian phylogenomic project data. GigaScience Database. http://dx.doi.org/10.5524/101000 RIS BibTeX Text

The evolutionary relationship of modern birds is one of the most challenging questions in systematic biology and has been debated for centuries. We proposed to rebuild the avian phylogenetic tree by using whole genome data, thus we have collected genomes of 48 bird species, representing 36 orders of bird class. The chicken, zebrafinch, and turkey genomes, which were sequenced in Sanger method, were collected from public domain. Another three genomes, pigeon, peregrine falcon, and duck, have been published during the development of this project.
The data posted here include the full genome assemblies of 42 bird species, the repeat and gene annotation produced by our new pipeline, 8295 1:1 syntenic orthologous genes, and the whole genome alignment data for all bird species. The detailed information of the published genomes can be accessed from their own publications. The 42 genomes first released here were sequenced and assembled with NGS technology in whole genome shotgun strategy. Using an homology-based method, we annotated 13000~17000 protein-coding genes in each avian genome.
So far as we known, the avian phylogenomic project is the biggest comparative genomics project to date. The unprecedented genomic data presented here will contribute to the downstream analyses in many fields, including phylogenetics, comparative genomics, neurology, development biology, etc.
Below are listed the links to all the individual species data used in this study. In addition, the entire dataset has been compressed into a single archive file for those who wish to retrieve the complete set.

Adelie Penguin - *Pygoscelis adeliae* - PYGAD - doi:10.5524/100006
American Crow - *Corvus brachyrhynchos* - CORBR - doi:10.5524/101008
American Flamingo - *Phoenicopterus ruber ruber* - PHORU - doi:10.5524/101035
Anna's Hummingbird - *Calypte anna* - CALAN - doi:10.5524/101004
Bald Eagle - *Haliaeetus leucocephalus* - HALLE - doi:10.5524/101040
Barn Owl - *Tyto alba* - TYTAL - doi:10.5524/101039
Bar-tailed Trogon - *Apaloderma vittatum* - APAVI - doi:10.5524/101016
Brown Mesite - *Mesitornis unicolor* - MESUN - doi:10.5524/101030
Budgerigar - *Melopsittacus undulatus* - MELUN - doi:10.5524/100059
Carmine Bee-eater - *Merops nubicus* - MERNU - doi:10.5524/101029

Figure 2. http://dx.doi.org/10.5524/101000 The avian phylogenomics project dataset stub in GigaDB.

Through our association with DataCite, each dataset in *Giga*DB is assigned a DOI that can be used as a standard citation in the reference section of a paper, improving access and use of these data in articles by the authors and other researchers.
In order to make NGS data interpretation as accessible as data generation, we have implemented “GigaGalaxy” (http://galaxy.cbiit.cuhk.edu.hk). We have ported the popular Short Oligonucleotide Analysis Package (SOAP http://soap.genomics.org.cn) as well as supporting tools such as Contiguator2 (http://contiguator.sourceforge.net) into the Galaxy framework, to provide seamless

NGS mapping, de novo assembly, NGS data format conversion and sequence alignment visualization. Our vision is to create an open publication, review and analysis environment by integrating GigaGalaxy into the publication platform at *GigaScience* and together with GigaDB, to help integrate data and analyses used in publications. We have begun this effort by re-implementing the data procedures described by Luo *et al*., (Luo *et al*., 2012) as Galaxy workflows so that they can be shared in a manner which can be visualized and executed in GigaGalaxy. We hope to revolutionize the publication model with the aim of executable publications, where data analyses can be reproduced and reused.

Acknowledgments

We would like to thank the following for their financial support

BGI Research, Shenzhen, China
China National GeneBank, China
Chinese University of Hong Kong, Hong Kong
BBSRC, UK
Natural Environment Research Council, UK

In addition to the authors, we would like acknowledge the works carried out by the following people; Alexandra Basford, Huayan Gao (GigaScience), Yong Zhang (BGI; China National Genebank) Shaoguang Liang (BGI-SZ), Alex Wong, Dennis Chan (BGI-HK), Tin-Lap Lee (CUHK), Qiong Luo, Senghong Wang, Yan Zhou (HKUST) and Mark Viant (Birmingham Uni).

References

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. (2012): Soapdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1(1):18. doi: 10.1186/2047-217X-1-18

The 3000 rice genomes project. (2014): The 3,000 Rice Genomes Project. GigaScience 3:7. doi:10.1186/2047-217X-3-7

Zhang, G; Li,B; Li,C; Gilbert,MTP; Jarvis,E; The Avian Genome Consortium; Wang,J. (2014): The avian phylogenomic project data. GigaScience Database. doi:10.5524/101000

The MicroScope platform:
from data integration to a rule-based system for massive and high-quality microbial genome annotation

Jonathan MERCIER, Alexandre RENAUX, Adrien JOSSO, Aurélie GENIN-LAJUS, E'Krame JACOBY, François LE FEVRE, Guillaume ALBINI, Claude SCARPELLI, Claudine MEDIGUE, David VALLENET

Direction des Sciences du Vivant, CEA, Institut de Génomique, Genoscope, France
CNRS-UMR8030, Evry, France
Université d'Evry Val d'Essonne, Evry, France

Abstract. The emergence of the Next Generation Sequencing (NGS) generates an incredible amount of genomes, whereas curation efforts to annotate them tend to decrease despite some community initiatives (*1*). To ease this manual process, we develop the MicroScope platform: an integrated environment for the annotation and exploration of microbial genomes (*2*). It is made of three major components:

1. a management system to store and organize biological knowledge in relational databases
2. a production system to organize and execute workflows
3. a visualization system for expert analyses and data curation through a Web interface.

Following the success of the platform, we are improving its throughput analysis to integrate an increasing number of genomes in a reasonable human time while maintaining a high quality of annotations. In this way, we initiate new methodological and technical developments on the MicroScope data management and production systems. A specific focus is given on the use of rule-based systems for the management of workflows and for the consistency evaluation of functional annotations that are performed automatically and then expertised by biologists.

The MicroScope data management system is made of several relational databases. The Prokaryotic Genome DataBase (PkGDB) gathers internal genomic data, human expertise and computational results. This central model is enriched by the integration of numerous public databases collecting different types of biological entities (e.g. genomes and genes from nucleic databanks, proteins from UniProt, metabolic data from ChEBI, Rhea, KEGG and MetaCyc). To support continuous data integration and reconciliation of these external resources, we designed the Galileo (*3*) application based on AndroMDA (*4*) and Play (*5*) frameworks. The Galileo model manages the integration of several releases of the same biological resource and ensures unicity of biological objects from different resources with the use of internal business keys based on their key properties. For instance, molecules are identified through their InChl signature, and reactions by a combination of stoichiometry information and molecule signatures. In comparison to other initiatives for biological data integration using Semantic Web, our main purpose is the unification of entities based on their common properties to define public methods and queries. These services are provided through REST API.

The MicroScope production system orchestrates about 25 workflows, which combine various bioinformatics software. The goal of this system is to keep analyses up to date according to the integration of new genomes, the updates of public databases and new software versions. Five years ago, we adopted the jBPM (*6*) framework to design our workflows. This framework allowed us to gain synchronization, robustness, control and traceability in the execution of million of jobs on HPC clusters (*7*). To increase the throughput of our analyses and the flexibility in the decision-making process of our production system, we will progressively switch to a new API called BIRDS (BioInformatics Rules Driven System) and developed at the Genoscope. BIRDS is based on the Drools framework (*8*) and provides a common environment for business rules and resource-driven workflows to automate bioinformatics treatments. This decision process integrated in a large data management system is an original feature of BIRDS in comparison to other workflow initiatives in biology (e.g. Taverna, Galaxy).

One important goal is to ease the human interpretation of genomic data in the light of predicted functions and biological processes (e.g. metabolic pathways). We are working on an explicit representation of the biological knowledge and on algorithmic tools designed to automate the biologist reasoning within the MicroScope platform. This application, named Grools, is a rule-based expert system. It is currently under development using the Drools framework. A first level of rules is designed to predict molecular functions according to protein domain composition and organism taxonomy. These rules were extracted and translated from the UniRule resource of UniProt database (*9*). The next step is to evaluate the overall coherence of these individual functions by applying logical rules between them and integrating additional information from biological processes where these functions may occur or not in a given organism. A first implementation of such deductive reasoning has been implemented in the HERBS system through a collaborative project between INRIA and SIB institutes (*10*). This can be applied, for example, crossing growth phenotypes on defined minimal media (e.g. Biolog phenotype microarray) and functional annotations to check the consistency of corresponding catabolic pathways.

In synergy with the technological progresses in the production system, biological data integration combined with logical reasoning should improve completeness and consistency of genome knowledge in the MicroScope platform. These IT innovations will be illustrated on the poster.

Keywords: Genome annotation, Data integration, Workflow, Rule-based system, Knowledge reasoning, Curation

References

1. R. Mazumder, D. A. Natale, J. A. E. Julio, L.-S. Yeh, C. H. Wu, Community annotation in biology. *Biol Direct* **5**, 12 (2010).
2. D. Vallenet, E. Belda, A. Calteau, S. Cruveiller, S. Engelen, A. Lajus, F. Le Fevre, C. Longin, D. Mornico, D. Roche, Z. Rouy, G. Salvignol, C. Scarpelli, A. A. Thil Smith, M. Weiman, C. Medigue, MicroScope–an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Res.* **41**, D636–647 (2013).
3. F. Le Fèvre, A. Josso, Galileo. galileo.genoscope.cns.fr (2014).
4. M. Bohlen *et al.*, AndroMDA. *AndroMDA*, www.andromda.org (2007).
5. G. Bort *et al.*, playframework. *Play framework*, www.playframework.com (2009).
6. T. Baeyens *et al.*, A java Business Process Management. *Java Business Process Management*, jbpm.jboss.org (2004).
7. D. Vallenet, S. Engelen, D. Mornico, S. Cruveiller, L. Fleury, A. Lajus, Z. Rouy, D. Roche, G. Salvignol, C. Scarpelli, C. Medigue, MicroScope: a platform for microbial genome annotation and comparative genomics. *Database (Oxford)* **2009**, bap021 (2009).
8. B. McWhirter *et al.*, Drools. drools.jboss.org (2001).
9. UniProt Consortium, Activities at the Universal Protein Resource (UniProt). *Nucleic acids research* **42**, D191–D198 (2014).
10. A. Viari, HERBS: A rule based system for checking the annotations of complete proteomes, Personal communication, (2014).

The e!DAL Java API: Sharing and Citing Research Data in Life Sciences

Daniel Arend*, Jinbo Chen, Christian Colmsee, Steffen Flemming, Uwe Scholz and Matthias Lange

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Corrensstr. 3,
OT Gatersleben, 06466 Stadt Seeland, Germany

{*corresponding author: arendd@ipk-gatersleben.de}

1 Background

The data life cycle from experiments to scientific publications follows in general the schema: experiments, data analysis, interpretation, and publication of scientific paper. Besides the publication of scientific findings, it is important to keep the data investment and ensure its future processing. This implies a guarantee for a long-term preservation and preventing of data loss. Condensed and enriched with metadata, primary data would be a more valuable resource than the re-extraction from articles. In this context, it becomes essential to change the handling and the acceptance of primary data within the scientific community. Data and publications should be honored with a high attention and reputation for data publishers.

2 The e!DAL data publication pipeline

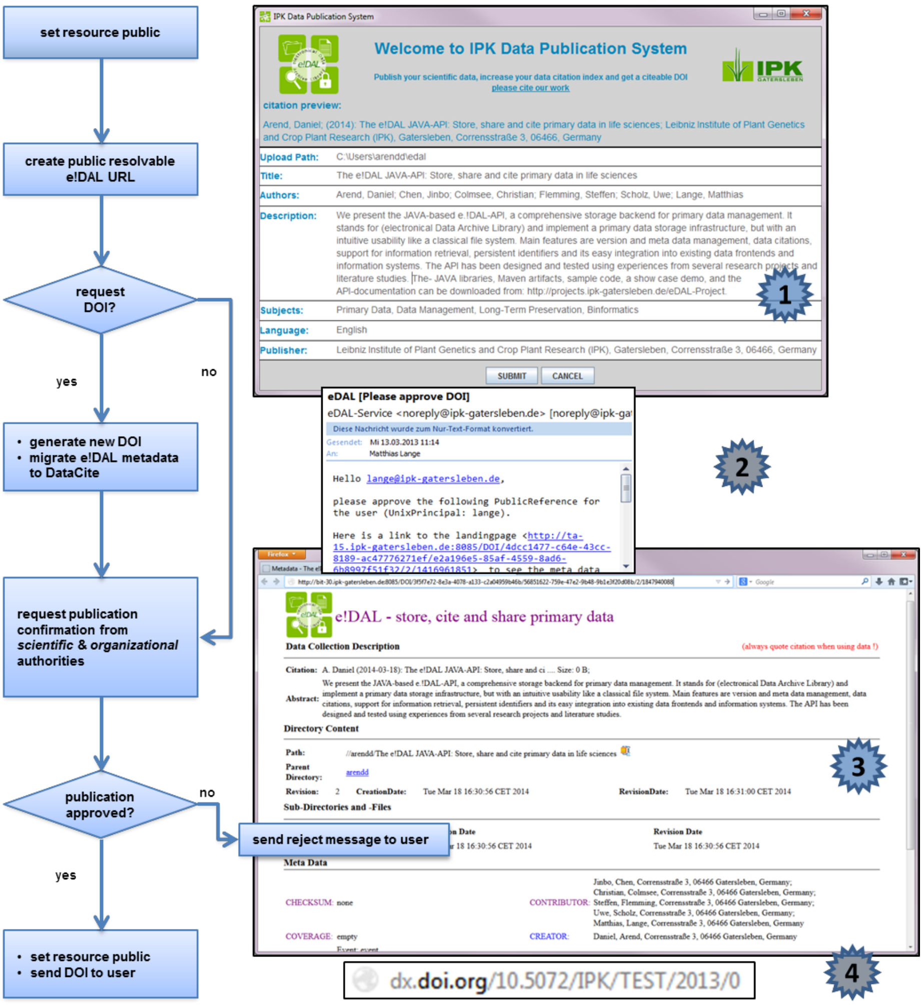
Here, we present new features of the e!DAL Java API (http://edal.ipk-gatersleben.de) as a lightweight software framework for publishing and sharing of research data [1]. e!DAL stands for electronic Data Archive Library. Its main features (Table 1) are version tracking, management of metadata, information retrieval, registration of persistent identifier, embedded HTTP(S) server for public data access, access as network file system, and a scalable storage backend. e!DAL is available as an open- source API for a local non-shared storage and remote usage to feature distributed applications.

Features	**Applied frameworks, standards and APIs**
Version Management – sequence of versions for the data set and its metadata	• H2 database • Hibernate • File system
Metadata Management – minimal set of technical and administrative metadata to ensure the mid-term data access of the stored data set	• DublinCore
Information Retrieval – search and retrieve relevant data sets for keyword queries over the metadata	• Apache Lucene • Hibernate Search • Apache SolR
Persistent Identifiers – provide persistent identifiers for an long-term stable public access of published data sets	• DOIs • URLs
Data Security – fine grained authorization of API methods, referred object and authenticated subjects (user)	• AspectJ • Java Authentications & Authorization API
Interoperability – seamless integration into existing infrastructures/tools	• local/remote Java API • GUI components • WebDAV • HTTP(S) Server

Table 1. Main features of e!DAL are listed in column one. For its implementation used frameworks, standards and APIs are referenced in column two.

The Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) is an approved data center belonging to the international DataCite consortium (http://www.datacite.org/) and applies e!DAL as data submission and registration system. In the latest version the focus was to extend the features for the registration of Digital Object Identifier (DOI) and the development of a simple, but sufficient approval process to regulate the assignment of persistent identifier. An intuitive publication tool (Figure 1), allows uploading your data into your own private repository over the web and getting a DOI to permanently reference the datasets and increase your "data citation" index.

The e!DAL approval process for data publication is based on an email notification system. After successfully proven a request for a DOI, the API automatically transfer all necessary files and metadata to DataCite and reply an email to the submitting user with his final assigned DOI. This ID permanently references to a virtual content page.


set resource public
create public resolvable e!DAL URL
request DOI?
yes
no
• generate new DOI
• migrate e!DAL metadata to DataCite
request publication confirmation from *scientific & organizational* authorities
publication approved?
no
yes
send reject message to user
• set resource public
• send DOI to user
1
2
3
4
SUBMIT
CANCEL

IPK Data Publication System

Welcome to IPK Data Publication System

Publish your scientific data, increase your data citation index and get a citeable DOI
please cite our work

citation preview:

Arend, Daniel; (2014): The e!DAL JAVA-API: Store, share and cite primary data in life sciences; Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Corrensstraße 3, 06466, Germany

Upload Path: C:\Users\arendd\edal

Title: The e!DAL JAVA-API: Store, share and cite primary data in life sciences

Authors: Arend, Daniel; Chen, Jinbo; Colmsee, Christian; Flemming, Steffen; Scholz, Uwe; Lange, Matthias

Description: We present the JAVA-based e.!DAL-API, a comprehensive storage backend for primary data management. It stands for (electronical Data Archive Library) and implement a primary data storage infrastructure, but with an intuitive usability like a classical file system. Main features are version and meta data management, data citations, support for information retrieval, persistent identifiers and its easy integration into existing data frontends and information systems. The API has been designed and tested using experiences from several research projects and literature studies. The- JAVA libraries, Maven artifacts, sample code, a show case demo, and the API-documentation can be downloaded from: http://projects.ipk-gatersleben.de/eDAL-Project.

Subjects: Primary Data, Data Management, Long-Term Preservation, Binformatics

Language: English

Publisher: Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Corrensstraße 3, 06466, Germany

eDAL [Please approve DOI]

eDAL-Service <noreply@ipk-gatersleben.de> [noreply@ipk-gat

Diese Nachricht wurde zum Nur-Text-Format konvertiert.

Gesendet: Mi 13.03.2013 11:14

An: Matthias Lange

```
Hello lange@ipk-gatersleben.de,

please approve the following PublicReference for
the user (UnixPrincipal: lange).

Here is a link to the landingpage <http://ta-
15.ipk-gatersleben.de:8085/DOI/4dcc1477-c64e-43cc-
8189-ac47776271ef/e2a196e5-85af-4559-8ad6-
6b8997f51f32/2/1416961851>  to see the meta data
```

http://bit-30.ipk-gatersleben.de:8085/DOI/3f5f7e72-8e3a-4078-a133-c2a04959b46b/56851622-759e-47e2-9b48-9b1e3f20d08b/2/1847940088

e!DAL - store, cite and share primary data

Data Collection Description (always quote citation when using data !)

Citation: A. Daniel (2014-03-18): The e!DAL JAVA-API: Store, share and ci .... Size: 0 B;

Abstract: We present the JAVA-based e.!DAL-API, a comprehensive storage backend for primary data management. It stands for (electronical Data Archive Library) and implement a primary data storage infrastructure, but with an intuitive usability like a classical file system. Main features are version and meta data management, data citations, support for information retrieval, persistent identifiers and its easy integration into existing data frontends and information systems. The API has been designed and tested using experiences from several research projects and literature studies.

Directory Content

Path: //arendd/The e!DAL JAVA-API: Store, share and cite primary data in life sciences

Parent Directory: arendd

Revision: 2 CreationDate: Tue Mar 18 16:30:56 CET 2014 RevisionDate: Tue Mar 18 16:31:00 CET 2014

Sub-Directories and -Files

n Date Revision Date

r 18 16:30:56 CET 2014 Tue Mar 18 16:30:56 CET 2014

Meta Data

CHECKSUM: none

CONTRIBUTOR: Jinbo, Chen, Corrensstraße 3, 06466 Gatersleben, Germany; Christian, Colmsee, Corrensstraße 3, 06466 Gatersleben, Germany; Steffen, Flemming, Corrensstraße 3, 06466 Gatersleben, Germany; Uwe, Scholz, Corrensstraße 3, 06466 Gatersleben, Germany; Matthias, Lange, Corrensstraße 3, 06466 Gatersleben, Germany

COVERAGE: empty

Event: event

CREATOR: Daniel, Arend, Corrensstraße 3, 06466 Gatersleben, Germany

dx.doi.org/10.5072/IPK/TEST/2013/0

Fig. 1. Schema of the e!DAL Publicaton workflow. e!DAL provides an easy usable publication interface for DOIs with a graphical user interface (1) and a simple approval process, which allows to define different reviewers, who can accept or revise every requested ID. The process is based on an email notification system (2). When the survey was successful, the API automatically transfer all necessary files and metadata to DataCite and the requesting user get an email with his final assigned DOI (4). This ID permanently references to a virtual content page (3), where you can download the files and check the corresponding metadata.

In addition we implement some new graphical components, like an easy installation/demo wizard, to simplify the deployment of a repositories using e!DAL.

3 Conclusion

e!DAL is a lightweight software framework for the management, publication, and sharing of research data. It is designed to turn sets of primary data into citable data publications. This is particularly important for the life sciences, where there is a big gap between the rate of data collection and the rate of data publication. Its well-defined API supports seamless integration into existing data-management software and infrastructures. In addition, e!DAL can be used as a supplement to manage primary data. Furthermore, its modular architecture and incorporated standards ensure version management, documentation, information retrieval, persistent identification, data security, and data publication. Developed within a context for the life sciences, e!DAL has many generic features that make it easily and readily applicable to other areas of science faced with similar needs. The e!DAL software is proven and has been deployed into the Maven Central Repository. Documentation and Software are also available at: http://edal.ipk-gatersleben.de.

4 References

[1] - D. Arend, M. Lange, C. Colmsee, S. Flemming, J. Chen and U. Scholz. *The e!DAL JAVA-API: Store, share and cite primary data in life sciences*. In IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2012, pages 1–5. 2012. DOI: 10.1109/BIBM.2012.6392737

Identifiers.org: integration tool for heterogeneous datasets

Camille Laibe[1], Sarala Wimalaratne[1], Nick Juty[1], Nicolas Le Novère[2], Henning Hermjakob[1]

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD

[2]Babraham Institute, Babraham Research Campus, Cambridge, CB22 3AT

Background

Generally, data providers identify individual records within their datasets using an identifier or accession number. Commonly, those identifiers are only unique within the dataset they originate from. For example, '9606' identifies "Homo sapiens" in the NCBI Taxonomy, but also identifies "Catha edulis" in the GRIN Taxonomy for Plants, the publication "Kohlhausen K. Das Offentliche Gesundheitswesen 1976 38(7):424-430" in PubMed, the "parathyroid hormone" in HGNC, "3-Fluorotoluene" in PubChem, etc. Additionally, the data is often distributed via multiple resources, using different URLs to access individual records. For example, the enzyme nomenclature information is available from at least 4 different sources: ExplorEnz (Trinity College, Dublin), KEGG Enzyme Database (Kyoto University Bioinformatics Center), ENZYME (Swiss Institute of Bioinformatics) and IntEnz (EMBL-European Bioinformatics Institute).

In order to address those issues related to the increased proliferation of ambiguous identifiers and non-perennial access URLs, an effort was launched in 2006 to provide a system through which appropriate URIs (Uniform Resource Identifiers) could be generated, based on existing local record identifiers already assigned by the data providers (http://identifiers.org/registry) [1].

A resolving system (http://identifiers.org/) [2], was launched to support requests from the Semantic Web community to provide these as HTTP URIs. These URIs are directly incorporable in datasets and usable by Semantic Web applications. Software tools handling data using those URIs need little work to process them and display them in a meaningful way to the end user (these URIs can actually be used as they stand in web interfaces). Moreover, these URIs are free, and provide unique, perennial and location-independent identifiers.

Here, the underlying Registry acts as the central storage repository for general information on the various datasets (termed *collections*), *namespace* information (unique short string identifying the collections), and lists the *resources* (physical locations from where data records can be retrieved).

The infrastructure is already used very successfully by, for example, the computational modelling community, which requires the ability to perennially record cross-references and links to external data records, despite the ever changing nature of the location of information on the web. It is being continually improved to meet the growing needs of new user communities.

Results

Identfiers.org is now running on multiple servers in 2 redundant data centres in London, which provide more reliable, robust and faster services to the community.

The semantics of the URIs handled by the resolver have been enhanced by having 2 distinct types of URIs: one for identifying the entity concept, and one for identifying the information recorded by the Registry about it. This means that canonical URIs (which uniquely identify data entities and are of the form http://identifiers.org/[namespace]/[entity]) provide more directly usable information for end users (previously a list all possible resources was provided, now one default resource is displayed, with a link to the former page). This default resource is selected using an algorithm which uses various properties such as resource reliability ('uptime') and data ownership (designated 'primary' for resources maintained directly by the data provider) status of the resources. Registry records of data entities are identified by URIs of the form http://info.identifiers.org/[namespace]/[entity]. Those URIs can be used to identify and retrieve metadata provided by the Registry or access information in various formats, such as RDF/XML (via content negotiation).

With the growing uptake of Linked Data in life sciences, exemplified by projects such as Bio2RDF [3] and the launch of the EBI RDF platform [4], new functionality and improvements to the infrastructure have been made. For example, there are now numerous URIs used to identify equivalent data records. These pose a challenge when integrating across heterogeneous datasets, even if all are encoded in RDF. In order to facilitate such integration, we have extended the services provided by Identifiers.org in order to serve a SPARQL compliant endpoint for URI schemes conversion. This allows the conversion of a URI from one scheme into a URI from a different one. To allow such functionality, the Registry records various URI scheme formats within the Registry, which are used to generate alternative URIs, using the Identifiers.org URIs as a canonical form. This allows the system to handle modification of the entity identifier, such converting http://identifiers.org/go/GO:0006915 into http://purl.obolibrary.org/obo/GO_0006915. Recording the alternative URI scheme formats requires additional curation efforts, but provides accurate results and should prevent any false positive.

With the growth of Linked Data and Semantic Web efforts, it has also become increasingly important not only to know where data may be accessed, but also the forms or formats in which it is available. Many resources now provide their datasets in a variety of formats on top of the standard HTML format. There are users who specifically require data encoded in a particular format, such as RDF/XML, or JSON. To allow direct data access to records encoded in specific formats, the Registry data model has been extended to enable it to record the various formats provided by a resource and how to access them. This information is stored at the level of the individual resources that offer records for a data collection, since each resource may offer different formats for consumption. Additionally, Identifiers.org has been updated to allow access to this information via content negotiation, performed on a user request.

Conclusion

Identifiers.org URIs are now a core element of numerous data management and provision infrastructures, such as the OpenPHACTS project [5] and the EBI RDF Platform. The extension of the information recorded in the underlying Registry, and the services provided, should greatly help in the integration of heterogeneous datasets. These incremental improvements to the infrastructure facilitate data integration independently of the URI scheme that may have been used originally, and allows the system to be more universally suitable for other use cases.

Acknowledgements

This work received support from the BBSRC (BB/K016946/1), IMI (OpenPhacts/115191) and EMBL.

References

1. Laibe, C., Le Novère, N.: MIRIAM Resources: tools to generate and resolve robust cross-references in Systems Biology. BMC Syst. Biol. 1:58 (2007)
2. Juty, N., Le Novère, N., Laibe, C.: Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Res. 40(D1):D580-D586 (2011)
3. Belleau, F., Nolin, M.A., Tourigny, N., Rigault, P., Morissette, J.: Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. J. Biomed. Inform. 41:706–716 (2008)
4. Jupp, S., Malone, J., Bolleman, J., Brandizi, M., Davies, M., Garcia, L., Gaulton, A., Gehant, S., Laibe, C., Redaschi, N., Wimalaratne, S.M., Martin, M., Le Novère, N., Parkinson, H., Birney, E., Jenkinson, A.M.: The EBI RDF platform: linked open data for the life sciences. Bioinformatics 1–2 (2014)
5. Gray, A., Groth, P., Loizou, A.: Applying linked data approaches to pharmacology: Architectural decisions and implementation. Semant. Web. 0:1–13 (2012)

Search QTL candidate genes by LAILAPS information retrieval system

Jinbo Chen*, Daniel Arend, Christian Colmsee, Maria Esch, Uwe Scholz, Matthias Lange

Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Corrensstr. 3, OT Gatersleben, 06466 Stadt Seeland, Germany

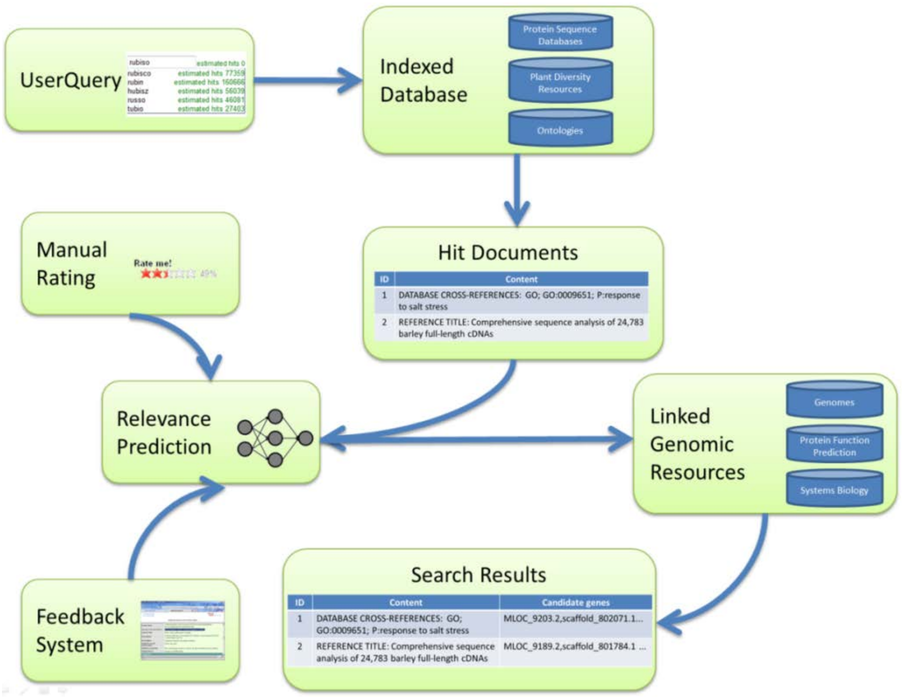
{* corresponding author: chenj@ipk-gatersleben.de}

1 Introduction

Correct identification of causative genes for an important agronomic trait is a very time-consuming task. Marker assisted breeding aims to identify relevant QTLs. But they often span genomic regions that can contain hundreds of candidate genes. The evaluation of potential functional candidates requires the integration of information from many different sources and its relevance ranking.

LAILAPS [1] is an integrative search engine in the genomics data domain. It links genome data to traits, i.e. protein functions, pathways, morphological phenotypes, agronomical and medical properties etc. LAILAPS implements a relevance ranking model that apply a machine learning approach to estimate the relevance of documents in public databases. The relevance depends on features of the matching documents and user relevance preferences. Among others such features are the commonly used hit frequency of search term in a document, hit surrounding keywords, the attribute context and number of outgoing links to other databases.

The first step in the LAILAPS query workflow (Figure 1) is a keyword based query, which will be enhanced by spelling correction and synonym expansion. Next step is query a full text index of about 55 million database records, which were extracted from 13 public databases (see http://lailaps.ipk-gatersleben.de/indexed_databases.html).


UserQuery
rubiso estimated hits 0
rubisco estimated hits 77359
rubin estimated hits 160666
hubisz estimated hits 56039
russo estimated hits 46081
tubio estimated hits 27403
Indexed Database
Protein Sequence Databases
Plant Diversity Resources
Ontologies
Hit Documents
Manual Rating
Rate me!
49%
Relevance Prediction
Linked Genomic Resources
Genomes
Protein Function Prediction
Systems Biology
Feedback System
Search Results

Hit Documents

ID	Content
1	DATABASE CROSS-REFERENCES: GO; GO:0009651; P:response to salt stress
2	REFERENCE TITLE: Comprehensive sequence analysis of 24,783 barley full-length cDNAs

Search Results

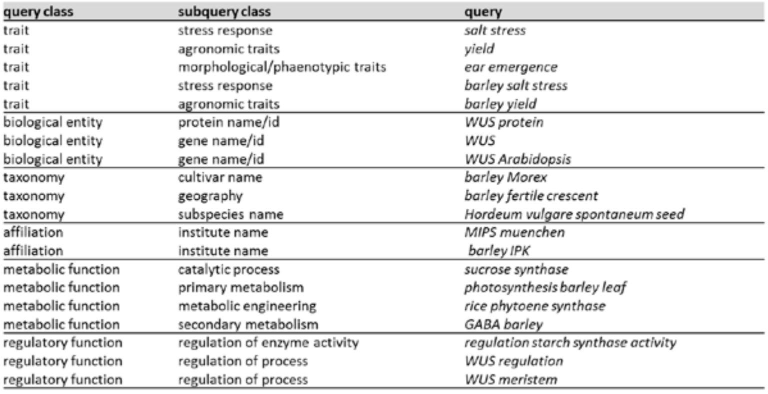
ID	Content	Candidate genes
1	DATABASE CROSS-REFERENCES: GO; GO:0009651; P:response to salt stress	MLOC_9203.2,scaffold_802071.1...
2	REFERENCE TITLE: Comprehensive sequence analysis of 24,783 barley full-length cDNAs	MLOC_9189.2,scaffold_801784.1 ...

Fig. 1. Schematic illustration of the LAILAPS query workflow

Fig. 2. The list of result database records is ranked by an artificial neural network. The training can be done explicitly using an expert curated reference set, implicitly by tracking user behaviour at the LAILAPS web site and estimated from her interaction pattern the relevance of the visited web page. For each of the ranked database records, LAILAPS computes linked genomic resources. The last step is the rendering the ranked results in LAILAPS Web frontend.

2 LAILAPS features at a glance

Machine learning-based ranking: Ranking is a central part of many information retrieval systems, such as document retrieval, collaborative filtering, computational advertising (online ad placement). A supervised machine learning ranking system was implemented in LAILAPS, which makes it possible to provide personalized search results. The order of the search hits are computed by an artificial intelligence driven relevance ranking, which has been trained by domain experts and evaluated for QTL candidate gene prediction (Table 1).



query class	subquery class	query
trait	stress response	*salt stress*
trait	agronomic traits	*yield*
trait	morphological/phaenotypic traits	*ear emergence*
trait	stress response	*barley salt stress*
trait	agronomic traits	*barley yield*
biological entity	protein name/id	*WUS protein*
biological entity	gene name/id	*WUS*
biological entity	gene name/id	*WUS Arabidopsis*
taxonomy	cultivar name	*barley Morex*
taxonomy	geography	*barley fertile crescent*
taxonomy	subspecies name	*Hordeum vulgare spontaneum seed*
affiliation	institute name	*MIPS muenchen*
affiliation	institute name	*barley IPK*
metabolic function	catalytic process	*sucrose synthase*
metabolic function	primary metabolism	*photosynthesis barley leaf*
metabolic function	metabolic engineering	*rice phytoene synthase*
metabolic function	secondary metabolism	*GABA barley*
regulatory function	regulation of enzyme activity	*regulation starch synthase activity*
regulatory function	regulation of process	*WUS regulation*
regulatory function	regulation of process	*WUS meristem*

Table 1. List of 20 traits and its expression as keyword queries. From the query results a set of randomly selected database entries were rated in respect to five relevance classes “fully agree”, ”minor quality doubts”, “could be of relevance”, “undecided”, and “no relevance”. The result of this curation is a set of 400 relevance ranked data base entries, which is available as supplement material [3].

As result of this expert curation, LAILAPS neural network was trained. Its performance was compared to the Apache Lucene [2] relevance scoring, which is one of the most prominently used text search API for relevance ranking in life science information systems and databases. Figure 2 show LAILAPS connect based feature model compared to the Lucene term frequency – inverse document frequency (TF/IDF) scoring.

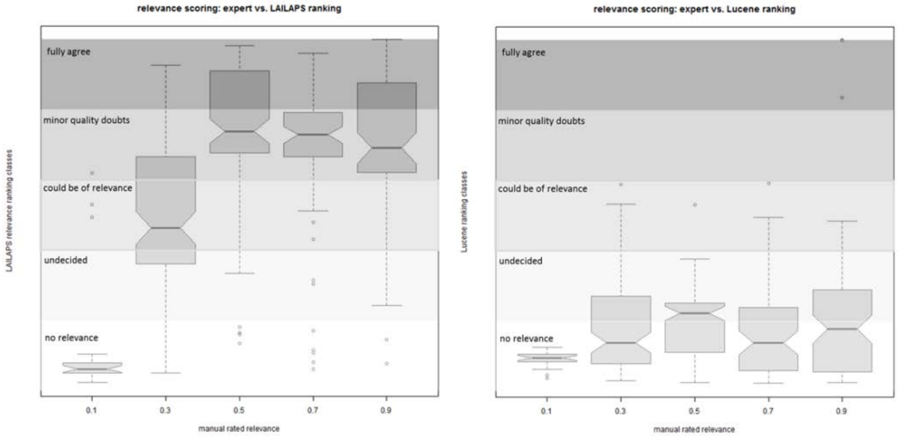


Fig. 3. Compare with Lucene Ranking method, LAILAPS ranking system can draw a clear separation between non-relevance results and relevance results, but weak for separation „very good“ from „good”

Built-in recommender system: To obtain similar data records, a popular and in IR well established method was used. First, getting interesting terms from a given document which mean terms whose TF-IDF score is high, and then searching the documents that contain these interesting terms.

Integration into 3rd-party websites and Drupal CMS: LAILAPS portlet was designed to embed the search engine into 3rd-party web pages (Figure 3). Parameters enable the customizing of service end point, database filtering in query result as well as the layout in the client web page. The embedded HTML query field provides all standard LAILAPS features like "query suggestion" and result estimation. Furthermore, content management systems (CMS) are supported. For example a Drupal module was developed that wraps the LAILAPS portlet and integrate it into the open source CMS Drupal.

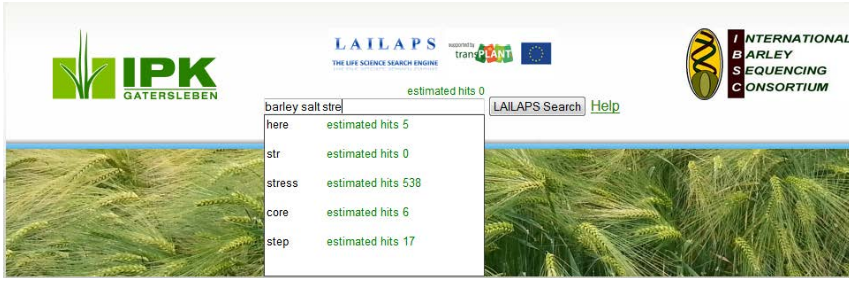
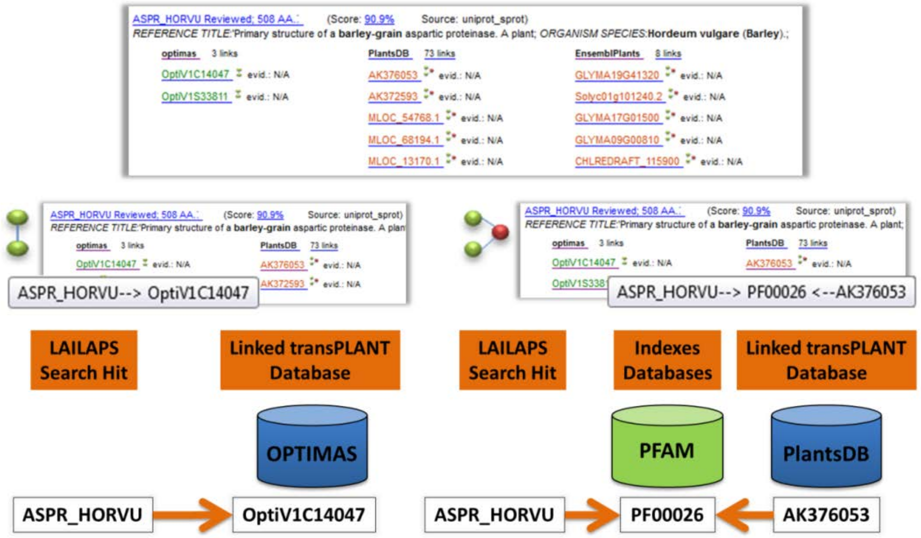


Fig. 4. LAILAPS portlet is integrated into web site of the International Barley Sequencing Consortium (`http://barleysequence.org/`)

Query suggestions: We present the design of a multi-stage query suggestion workflow and its implementation in the life science IR system LAILAPS. The workflow includes enhanced tokenisation, word breaking, spelling correction, query expansion and query suggestion ranking.

Link traits to genomes: LAILAPS supports the integrative search over distributed data resources. Rather than integrate them tightly, we assume that genomics resources build a network. This motivates the idea of database integration by a reverse tracing of genome annotations. Basis of this approach are sequence annotation, like homology search, sequence pattern search or even manual curated annotation to literature databases. In sum, LAILAPS imported 80 million links to 13 databases and enables a reverse identifier lookup to link traits to the genomics resources (Figure 4).



ASPR_HORVU Reviewed; 508 AA.; (Score: 90.9% Source: uniprot_sprot)
REFERENCE TITLE:'Primary structure of a barley-grain aspartic proteinase. A plant; ORGANISM SPECIES:Hordeum vulgare (Barley).;

optimas 3 links	PlantsDB 73 links	EnsemblPlants 8 links
OptiV1C14047 evid.: N/A	AK376053 evid.: N/A	GLYMA19G41320 evid.: N/A
OptiV1S33811 evid.: N/A	AK372593 evid.: N/A	Solyc01g101240.2 evid.: N/A
	MLOC_54768.1 evid.: N/A	GLYMA17G01500 evid.: N/A
	MLOC_68194.1 evid.: N/A	GLYMA09G00810 evid.: N/A
	MLOC_13170.1 evid.: N/A	CHLREDRAFT_115900 evid.: N/A

Fig. 5. Example workflow to query candidate genes for the trait “barley grain”. The best ranked result has 84 referenced genes. For example “OptiV1C14047” in OPTIMAS database (`http://www.optimas-bioenergy.org/optimas_dw`) and “AK376053” in PlantsDB (`http://mips.helmholtz-muenchen.de/plant/barley/`). The link from “ASPR_HORVU” to “OptiV1C14047” is a “direct link”. Contrarily, there is no direct link (annotation) from “ASPR_HORVU” to “AK376053”. But “ASPR_HORVU” links to “PF00026” and “PF00026” links transitively to “AK376053”. We call this kind of link is “indirect link”.

Acknowledgements

This work was made in the frame of the transPLANT project and is funded by the European Commission within its 7th Framework Programme, contract number 283496.

References

1. M. Lange, K. Spies, C. Colmsee, S. Flemming, M. Klapperstück and U. Scholz. The LAILAPS Search Engine: A Feature Model for Relevance Ranking in Life Science Databases. Journal of Integrative Bioinformatics, 7(3):e118, 2010. `http://lailaps.ipk-gatersleben.de`
2. Apache Lucene. `http://lucene.apache.org/core`
3. M. Esch, Jinbo Chen and Matthias Lange: Training dataset of LAILAPS ranking system doi:10.5447/IPK/2014/2

PRIDE Proteomes: a Condensed View of the Plethora of Public Proteomics Data Available in the PRIDE Repository

Noemi del Toro, Florian Reisinger, Joseph M. Foster, Javier Contell, Antonio Fabregat, Pau Ruiz Safont, Henning Hermjakob and Juan Antonio Vizcaíno

European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom.

Abstract. The PRIDE (PRoteomics IDEntifications) database [1] (http://www.ebi.ac.uk/pride) at the European Bioinformatics Institute (EBI, Cambridge, UK) is one of the main public repositories for mass spectrometry (MS)-based proteomics data. In PRIDE, the amount of data is constantly growing at a very significant pace. Apart from the data management component, the other main challenge in proteomics resources such as PRIDE is to provide an aggregated and quality-scored version of the peptide/protein identifications found across all the submitted projects, in order to decide which information is more reliable.

PRIDE Proteomes is a new resource providing a condensed, protein centric view of the MS data in PRIDE. As part of the new project a scheme for quality scoring is being developed at present. The information that is used for this aim is the experimental metadata annotation, the number of evidences and the resulting data after applying the 'PRIDE Cluster' [2] algorithm. This scoring will be then propagated to the peptide and protein level by using a set of defined rules.

A beta version of PRIDE Proteomes is now available for four species (human, rat, mouse and *Arabidopsis*) at http://wwwdev.ebi.ac.uk/pride/proteomes.

Keywords: proteomics, bioinformatics, mass spectrometry, database, repository, protein, peptide, psm, pipeline, web service, web application

1 Introduction

The PRIDE database (http://www.ebi.ac.uk/pride) at the EBI is a public data repository for MS-derived proteomics data. It stores peptide and protein identifications, the corresponding mass spectra, post-translational modifications (PTMs), protein/peptide expression values (if available) and experimental metadata.

PRIDE is the initial submission point for MS/MS data in the ProteomeXchange consortium [3] (http://www.proteomexchange.org). ProteomeXchange provides an internationally co-ordinated repository infrastructure compatible with community requirements for data deposition and dissemination. Data is made accessible in ProteomeCentral (http://proteomecentral.proteomexchange.org).

The study-centric view of the PRIDE repository does not provide the opportunity to analyse and compare the information available across all the public data submitted. The main purpose of the PRIDE Proteomes project is to provide a homogeneous and integrated view of the data stored in PRIDE together with a metric that allows the user to be confident in the reliability of the data.

The main challenge is how to determine the quality of the proteomics identifications in a highly heterogeneous resource like PRIDE. This still remains an issue in the field since results from different search engines, especially in different experimental settings, are not directly comparable. Our approach to the problem was the development of the PRIDE Cluster algorithm [2], based on spectral clustering. Data quality assessment is performed at different levels: experiment (based on metadata annotation), protein, peptide and peptide-spectrum match (PSM).

3 System Architecture

The PRIDE Proteomes project is structured in three different layers: database and pipeline, web service and web application. All the layers have been developed in the Java language and provide different levels of abstraction and methods to access the data.

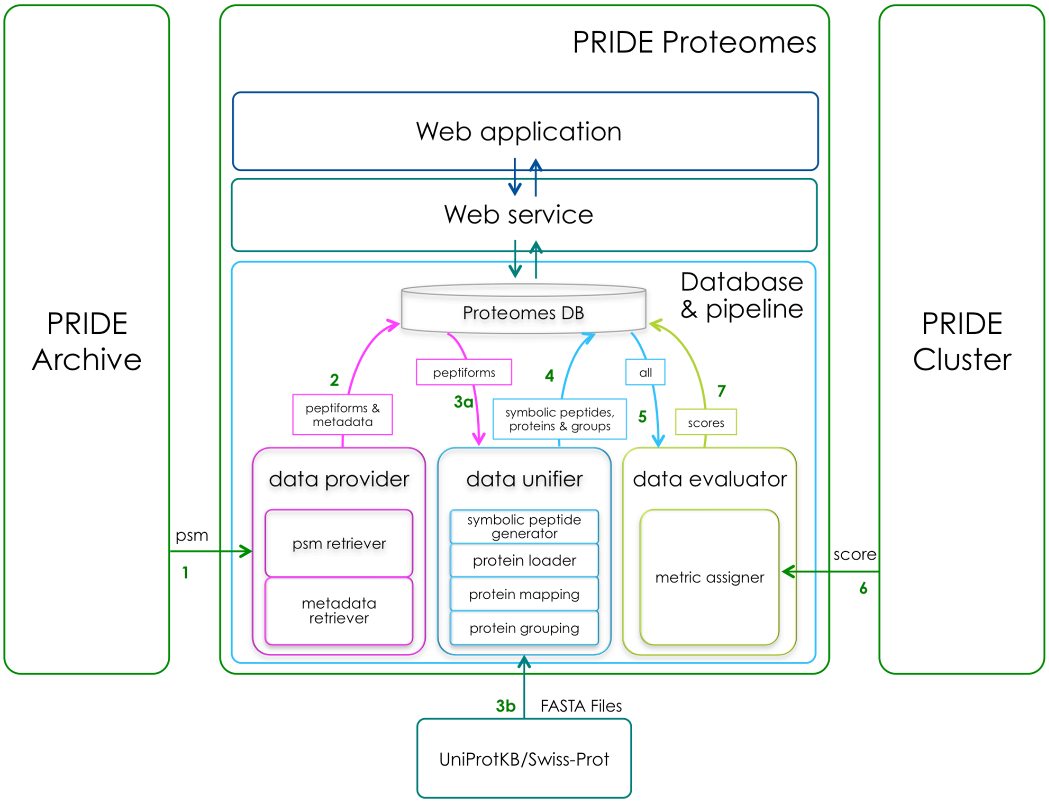


Fig. 1. Data flow of the PRIDE Proteomes project, from the PRIDE Archive, through the pipeline, to the web application available for the user.

3.1 Database & Pipeline

The database is the core of the PRIDE Proteomes project. Apart from the obvious role as a data store, the database is used as a means of communication between the different stages of the pipeline. The proteomes pipeline has been implemented with the Java Spring Framework. Inside the pipeline three sub-pipelines have been differentiated: data provider, data unifier and data evaluator. At the time of writing, the data provider and the data evaluator are still in early stages of development.

Data provider. The main goal of this sub-pipeline (Fig 1, steps 1 and 2) is to generate the peptides that represent the biological entity observed by MS experiments as the result of combining all the PSMs (Peptide Spectrum Matches) that share the same amino acid sequence and modifications (or their absence).

The PSMs are retrieved from the PRIDE Archive repository for the four species initially selected (human, rat, mouse and *Arabidopsis*) and are filtered by at least two criteria: the length of the peptide (between 6 and 100 amino acids) and all the amino acid letters that represent the sequence need to be unambiguous. Once the grouping stage is finished only the peptides created (called 'peptiforms', they represent a raw peptide sequence plus their modifications per species) are stored in the database (Fig 1, step 2).

Together with the peptiforms, the metadata from the projects associated with the PSMs are inserted in the database (Fig 1, step 2).

Data unifier. The pipeline behind the data unifier will be in charge of the next steps:

- Iterate through all the peptiforms generated in the previous stage (Fig. 1, step 3a) and extract from all of them the distinct peptide sequences (this set of sequences will be called 'symbolic peptides').
- In parallel to this process (Fig. 1, step 3b), the reference proteome is loaded into the database. For this purpose the corresponding UniProtKB FASTA files are downloaded, parsed and stored in the Proteomes database (Fig. 1, step 3).
- Once the proteins and the distinct sequences of the peptides are available in the database, an algorithm to map the peptide sequences to the protein sequences is executed. The corresponding start and end positions for the peptide in the protein sequence will be stored in the database, together with flags indicating whether the peptide is fully tryptic, and whether it is unique in the reference proteome (Fig. 1. step 4). .
- The last data unifier step allows the definition and generation of groups of proteins. By means of a simple comma-separated value (CSV) file, the pipeline can be configured with a specific criterion like the pertinence of the same gene and/or family of proteins to create the protein groups. After establishing the protein groups, the unique peptides for a group will be calculated and persisted. (Fig. 2, right, shows a representation of the groups in the web application).

Data evaluator. When all the information has been generated, the data evaluator phase starts. The evaluation of the quality of the evidences studied (Fig 1, steps 6 and 7) is propagated from the raw peptides or peptiforms, to the groups of proteins. Initially the peptiforms are associated a first value, or ranking, generated from the assessment of the PSMs that have helped to build them, and the associated metadata. This low level of quality checking has been implemented in previous prototypes of PRIDE Proteomes as a weighted function of the level of pertinence to one or several generated clusters from the 'PRIDE Cluster' algorithm [2]. In the future, reprocessed data by third parties will also be integrated.

3.2 Web Service

A RESTful web service (http://wwwdev.ebi.ac.uk/pride/ws/proteomes) exposes the content generated by the pipeline. It provides the communication between the data layer and the presentation layer enabling the decoupling among these components and provides a programmatic access for third parties.

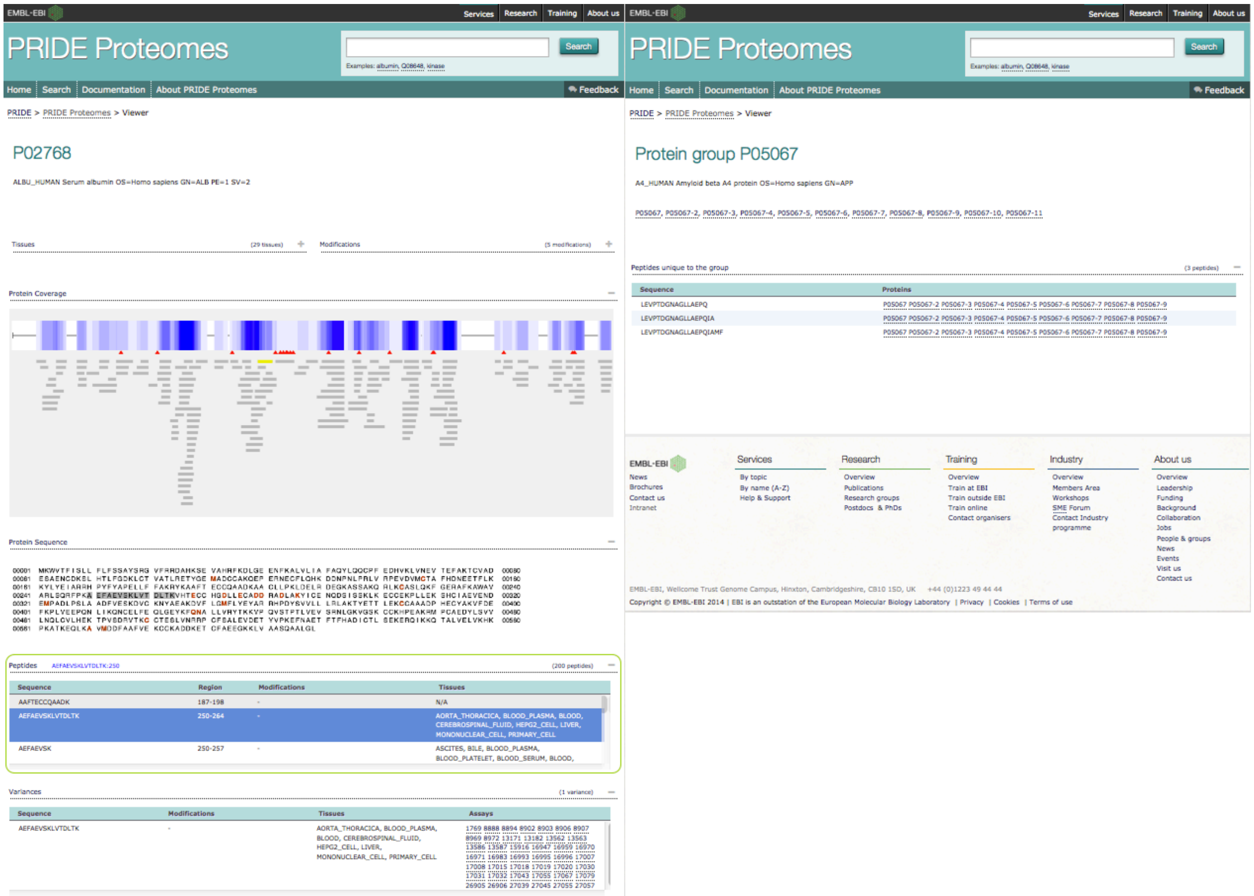
Through this service users can perform several queries to the system to retrieve the peptides by specific metadata associated. It is possible to find out if specific peptides are unique across all of the chosen reference proteomes, which can be at the group or protein level. A summary of the main web service methods can be found in on-line documentation provided for the web service.

3.3 Web Application

The PRIDE Proteomes web application provides a rich and highly interactive web interface to analyse, in the context of the protein, the location of the peptides and the PTMs. Complementing the protein-centric view, a 'protein group' view is also available to simplify the comparison between different related proteins (e.g. products of the same gene). In Fig. 2 a screenshot of the current version of the web application is shown.

4 Future Work

PRIDE Proteomes is still in development; the main foundations and core functionality have been established. In the next iteration circle the highest priority for the project will be to achieve a stable confidence metric.



P02768

ALBU_HUMAN Serum albumin OS=Homo sapiens GN=ALB PE=1 SV=2

Tissues (29 tissues) +

Modifications (5 modifications) +

Protein Coverage

Protein Sequence

```
00001  MKWVTFISLL FLFSSAYSRG VFRRDAHKSE VAHRFKDLGE ENFKALVLIA FAQYLQQCPF EDHVKLVNEV TEFAKTCVAD  00080
00081  ESAENCDKSL HTLFGDKLCT VATLRETYGE MADCCAKQEP ERNECFLQHK DDNPNLPRLV RPEVDVMCTA FHDNEETFLK  00160
00161  KYLYEIARRH PYFYAPELLF FAKRYKAAFT ECCQAADKAA CLLPKLDELR DEGKASSAKQ RLKCASLQKF GERAFKAWAV  00240
00241  ARLSQRFPKA EFAEVSKLVT DLTKVHTECC HGDLLECADD RADLAKYICE NQDSISSKLK ECCEKPLLEK SHCIAEVEND  00320
00321  EMPADLPSLA ADFVESKDVC KNYAEAKDVF LGMFLYEYAR RHPDYSVVLL LRLAKTYETT LEKCCAAADP HECYAKVFDE  00400
00401  FKPLVEEPQN LIKQNCELFE QLGEYKFQNA LLVRYTKKVP QVSTPTLVEV SRNLGKVGSK CCKHPEAKRM PCAEDYLSVV  00480
00481  LNQLCVLHEK TPVSDRVTKC CTESLVNRRP CFSALEVDET YVPKEFNAET FTFHADICTL SEKERQIKKQ TALVELVKHK  00560
00561  PKATKEQLKA VMDDFAAFVE KCCKADDKET CFAEEGKKLV AASQAALGL
```

Peptides AEFAEVSKLVTDLTK:250 (200 peptides)

Sequence	Region	Modifications	Tissues
AAFTECCQAADK	187-198	-	N/A
AEFAEVSKLVTDLTK	250-264	-	AORTA_THORACICA, BLOOD_PLASMA, BLOOD, CEREBROSPINAL_FLUID, HEPG2_CELL, LIVER, MONONUCLEAR_CELL, PRIMARY_CELL
AEFAEVSK	250-257	-	ASCITES, BILE, BLOOD_PLASMA, BLOOD_PLATELET, BLOOD_SERUM, BLOOD,

Variances (1 variance)

Sequence	Modifications	Tissues	Assays
AEFAEVSKLVTDLTK	-	AORTA_THORACICA, BLOOD_PLASMA, BLOOD, CEREBROSPINAL_FLUID, HEPG2_CELL, LIVER, MONONUCLEAR_CELL, PRIMARY_CELL	1769 8888 8894 8902 8903 8906 8907 8969 8972 13171 13182 13562 13563 13586 13587 15916 16947 16959 16970 16971 16983 16993 16995 16996 17007 17008 17015 17018 17019 17020 17030 17031 17032 17043 17055 17067 17079 26905 26906 27039 27045 27055 27057

Protein group P05067

A4_HUMAN Amyloid beta A4 protein OS=Homo sapiens GN=APP

P05067, P05067-2, P05067-3, P05067-4, P05067-5, P05067-6, P05067-7, P05067-8, P05067-9, P05067-10, P05067-11

Peptides unique to the group (3 peptides)

Sequence	Proteins
LEVPTDGNAGLLAEPQ	P05067 P05067-2 P05067-3 P05067-4 P05067-5 P05067-6 P05067-7 P05067-8 P05067-9
LEVPTDGNAGLLAEPQIA	P05067 P05067-2 P05067-3 P05067-4 P05067-5 P05067-6 P05067-7 P05067-8 P05067-9
LEVPTDGNAGLLAEPQIAMF	P05067 P05067-2 P05067-3 P05067-4 P05067-5 P05067-6 P05067-7 P05067-8 P05067-9

EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK +44 (0)1223 49 44 44

Copyright © EMBL-EBI 2014 | EBI is an outstation of the European Molecular Biology Laboratory | Privacy | Cookies | Terms of use

Fig. 2. Protein view (left) and protein group view (right) from the PRIDE Proteomes web application

5 Acknowledgments

We want to acknowledge funding from The Wellcome Trust [grant numbers WT085949MA and WT101477MA].

6 References

1. Vizcaíno J.A., Côté R.G., Csordas A., Dianes J.A., Fabregat A., Foster J.M., Griss J., Alpi E., Birim M., Contell J., O'Kelly G., Schoenegger A., Ovelleiro D., Pérez-Riverol Y., Reisinger F., Ríos D., Wang R. q, Hermjakob H. (2013) The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res* Volume 41 p.d1063-9
2. Griss, J., Foster, J. M., Hermjakob, H., and Vizcaino, J. A. (2013) PRIDE Cluster: building a consensus of proteomics data. *Nat Methods* 10, 95-96.
3. Vizcaíno J.A., Deutsch EW2, Wang R., Csordas A., Reisinger F., Ríos D., Dianes J.A., Sun Z., Farrah T., Bandeira N., Binz P.A., Xenarios I., Eisenacher M., Mayer G., Gatto L., Campos A., Chalkley R.J., Kraus H.J., Albar J.P., Martinez-Bartolomé S., Apweiler R., Omenn G.S., Martens L., Jones A.R., Hermjakob H. (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnology* 32, 223–226

A Machine Learning based Natural Language Interface for a database of medicines

Ricardo Ferrão, Helena Galhardas, Luísa Coheur

INESC-ID and Instituto Superior Técnico, Universidade de Lisboa, Portugal
ricardo.ferrao@tecnico.ulisboa.pt
helena.galhardas@tecnico.ulisboa.pt
luisa.coheur@inesc-id.pt

Abstract. Medical information needs to be efficiently retrieved by medical staff and common users. Most retrieval engines rely on search methods based on keywords or interfaces that enable to navigate through the index of a book or a website. MedicineAsk is a prototype that enables users to access information about medicines through a Natural Language Interface (NLI) in Portuguese. In this paper, we compare the existing rule-based NLI module of MedicineAsk against a machine learning based one. We show that the machine learning based NLI is a better alternative to the rule-based methods for most scenarios. This suggests the possibility of a hybrid technique to take advantage of both methods.

1 Introduction

Many interfaces to medical information require the user to know how this information is organized and, sometimes, to be a medicine expert. For instance, the Portuguese Infarmed website offers access to data about medicines through the *Prontuário Terapêutico* (Therapeutic Handbook) [1] either by navigating through an index or by keyword search.

A Natural Language Interface (NLI) is an alternative to this type of interfaces. In previous work we proposed the MedicineAsk prototype [3], which provides a NLI to the Infarmed website. MedicineAsk extracts information from this site and stores it in a relational database. Then, it is able to answer user's questions in Portuguese. Preliminary experiments showed that the MedicineAsk NLI provided a better user satisfaction and ease of use when compared to the Infarmed website [5].

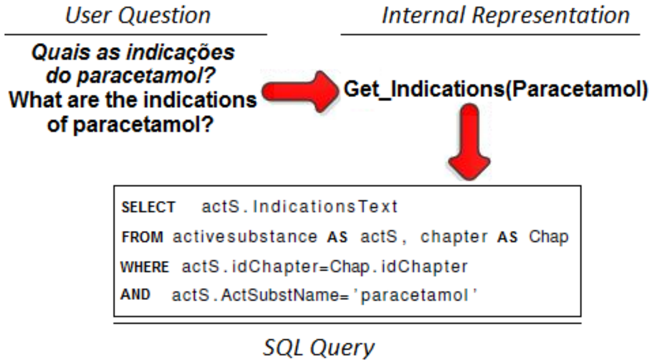
The NLI module of MedicineAsk is rule-based. In this paper, we follow a machine learning approach.

2 MedicineAsk: rule-based vs machine learning approach

Given a user question in Portuguese, the NLI module of MedicineAsk interprets and translates it into an SQL query, which is posed to the relational database that provides the answer. Figure 1 shows an example question and its internal representation

[1] From this point forward we shall refer to Infarmed's *Prontuário Terapêutico* as "the Infarmed website".

in MedicineAsk. The current interpretation step relies on handcrafted rules and keyword spotting [4]. Each user question is tested against each rule and, if a match occurs, the question type is identified (for instance, a question about indications represents a question type and a question about adverse reactions represents another question type). Moreover, a dictionary-based named entity recognizer is responsible for extracting medical entities. When the question type and the question entities are obtained, another set of rules generates the corresponding SQL query. If no rule can be applied to a given question, a keyword spotting technique is used.


User Question
Quais as indicações do paracetamol?
What are the indications of paracetamol?
Internal Representation
Get_Indications(Paracetamol)
SQL Query

```
SELECT  actS.IndicationsText
FROM activesubstance AS actS, chapter AS Chap
WHERE actS.idChapter=Chap.idChapter
AND  actS.ActSubstName='paracetamol'
```

Fig. 1: Example of a user question and its internal representations in MedicineAsk

In this paper, we use Support Vector Machines (SVM) to find the question type in a one-versus-all strategy, as they led to state-of-art results in question classification [7]. To this end, we use LUP [6], a platform for Natural Language Understanding that includes the LIBSVM[2] implementation of SVMs, and supports several features, such as unigrams, bigrams and word shape.

3 Experiments

The training corpus was built from 450 questions previously collected [5]. These questions were divided into 15 question types. We collected a test set to compare the rule-based with the machine learning approach. To this end, an on-line questionnaire composed of 9 different scenarios was distributed over the internet. Each scenario consists of a description of a problem that is related to medicines (e.g. “John needs to know the adverse reactions of Efferalgan, what kind of question should he ask?”). The 58 participants were invited to propose one or more (natural language) questions for each scenario. In this preliminary experiment we used questions from 30 randomly chosen users, which included a total of 296 questions divided into 9 scenarios. We tested these questions against a rule- and a SVM-based NLI. Figure 2 shows the percentage

[2] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/.

of questions correctly classified, for each scenario (*1 to 9*) and for the the average of all scenarios (*Total*). For each scenario, we show the percentage of questions correctly classified for each feature set used by SVM and for the rule-based NLI.

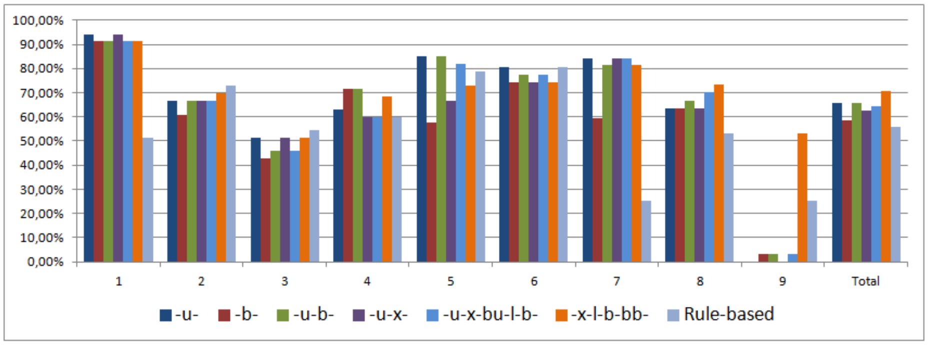


Fig. 2: Percentage of correctly classified questions by scenario for all users. The features are as follows: u - unigram, b - bigram, x - word shape, l - length, bu - binary unigrams, bb - binary bigrams.

Observing the total scores over all scenarios, we conclude that SVM has an advantage over the original rule-based method. This is due to machine learning methods being more flexible than rule-based ones. We can see that simply using unigrams can lead to very good results in most scenarios.

The majority of the cases in which the SVM failed were due to the fact that some words in the user's requests were not present in the corpus. For example, in the question *"Quais as doses pediátricas recomendadas da mizolastina?"* ("What are the recommended dosages for mizolastina?") we find that "doses", "pediátricas" and "recomendadas" are not present in the corpus. Also, some words were more frequently associated (in the training corpus) with a category different from the correct one. Scenario 9, which had the longest questions, got the worst results. Finally, none of the methods is robust to errors made by the user. For example, in some instances, the user misspelled certain words like medicine names.

4 Related Work

MEANS [2] is a question-answering system in the medical domain. Analogously to MedicineAsk, it creates a database, processes a question in natural language (in English), builds a query from that question and obtains the answer from the database. An ontology was defined in order to represent the concepts and relations used in MEANS. The database used is an RDF graph and the language used to query it is SPARQL. The database was created by extracting information from a medical corpus and annotating it in RDF. MEANS analyses questions through rule-based and machine learning methods.

It uses manually built patterns to determine the question type. To recognize named entities, the rule-based methods use MetaMap [1], an online tool which finds and classifies concepts in text by mapping them to concepts from the Unified Medical Language System (UMLS). The machine learning method uses Conditional Random Fields (CRFs) to classify the medical concepts. MEANS supports questions about general medicine unlike MedicineAsk which focuses on questions regarding medicines.

5 Conclusions

The overall results of these preliminary tests show that SVM outperforms the rule-based methods. Currently, MedicineAsk tries to answer a question through a rule-based method, and if that fails it relies on keywords. We intend to replace the keyword method by an SVM and perform experiments to validate the gain obtained. We intend to enrich the corpus to further improve SVM's results. We also consider to use an ontology to represent named medical entities and other machine learning algorithms (e.g., CRFs) to find them. We also consider to use Portuguese morphosyntactical features or other linguistically motivated features to improve the recognition of questions.

ACKNOWLEDGEMENTS

This work was partially supported by national funds through FCT - *Fundação para a Ciência e a Tecnologia*, under the project PEst-OE/EEI/LA0021/2013 and the DataStorm Research Line of Excellency funding (EXCL/EEI-ESS/0257/2012).

References

1. Alan R. Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: the metamap program. *AMIA Annu Symp Proc*, pages 17–21, 2001.
2. Asma Ben Abacha and Pierre Zweigenbaum. Medical question answering: Translating medical questions into SPARQL queries. *ACM SIGHIT International Health Informatics Symposium (IHI 2012)*, 2012.
3. Helena Galhardas, Vasco Duarte Mendes, and Luísa Coheur. Medicine.ask: a natural language search system for medicine information. *INFORUM 2012 - Simpósio de Informática*, 2012.
4. C. Jacquemin. *Spotting and discovering terms through natural language processing.* The MIT Press, 2001.
5. Vasco Duarte Mendes. Medicine.ask: an extraction and search system for medicine information. Master's thesis, Instituto Superior Técnico, 2011.
6. Pedro Mota, Lusa Coheur, Srgio dos Santos Lopes Curto, and Pedro Fialho. Natural language understanding: From laboratory predictions to real interactions. In *15th International Conference on Text, Speech and Dialogue (TSD)*, volume 7499 of *Lecture Notes in Artificial Intelligence*. Springer, September 2012.
7. J. Silva, L. Coheur, A. Mendes, and A. Wichert. From symbolic to sub-symbolic information in question classification. *Artificial Intelligence Review*, 2011.

AgreementMakerLight: A Scalable Automated Ontology Matching System

Daniel Faria[1], Catia Pesquita[1,2], Emanuel Santos[2], Isabel F. Cruz[3], and Francisco M. Couto[1,2]

[1] LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
[2] Dept. Informática, Faculdade de Ciências, Universidade de Lisboa, Portugal
[3] ADVIS Lab, Dept. of Computer Science, University of Illinois at Chicago, USA

Abstract. Ontology matching is a critical task to enable interoperability between the numerous life sciences ontologies with overlapping domains. However, it is a task made difficult by the size of many of these ontologies.

AgreementMakerLight (AML) is a scalable automated ontology matching system developed primarily for the life sciences domain. It can handle large ontologies efficiently, specializes in the use of background knowledge, includes an innovative alignment repair algorithm, and features a graphical user interface which makes it easy to use.

AML has obtained top results in matching life sciences ontologies in the Ontology Alignment Evaluation Initiative, and is being used in several other applications.

1 Background

Ontology matching is the task of finding correspondences (or mappings) between semantically related concepts of two ontologies, so as to generate an alignment that enables integration and interoperability between those ontologies [2]. This task is particularly relevant in the life sciences, given the boom in ontology development which gave rise to hundreds of life sciences ontologies with partially overlapping domains.

At its base, ontology matching is a problem of quadratic complexity as it entails comparing all concepts of one ontology with all concepts of the other. Early ontology matching systems were not overly concerned with scalability, as the matching problems they tackled were relatively small. But with the increasing interest in matching large biomedical ontologies, scalability became a critical aspect, and the traditional all-versus-all ontology matching strategy became unfeasible.

AgreementMakerLight (AML) is a scalable automated ontology matching system developed to tackle large ontology matching problems, and focused in particular on the biomedical domain. It is derived from AgreementMaker, one of the leading first generation ontology matching systems [1].

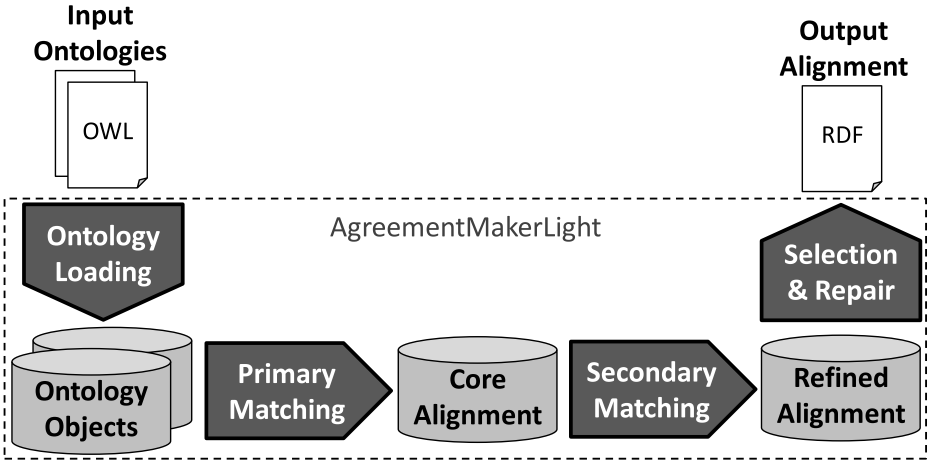


Fig. 1. AgreementMakerLight ontology matching framework.

2 The AgreementMakerLight System

AML is an open source system that is available through github [1] both as a runnable Jar and as an Eclipse project.

2.1 Ontology Matching Framework

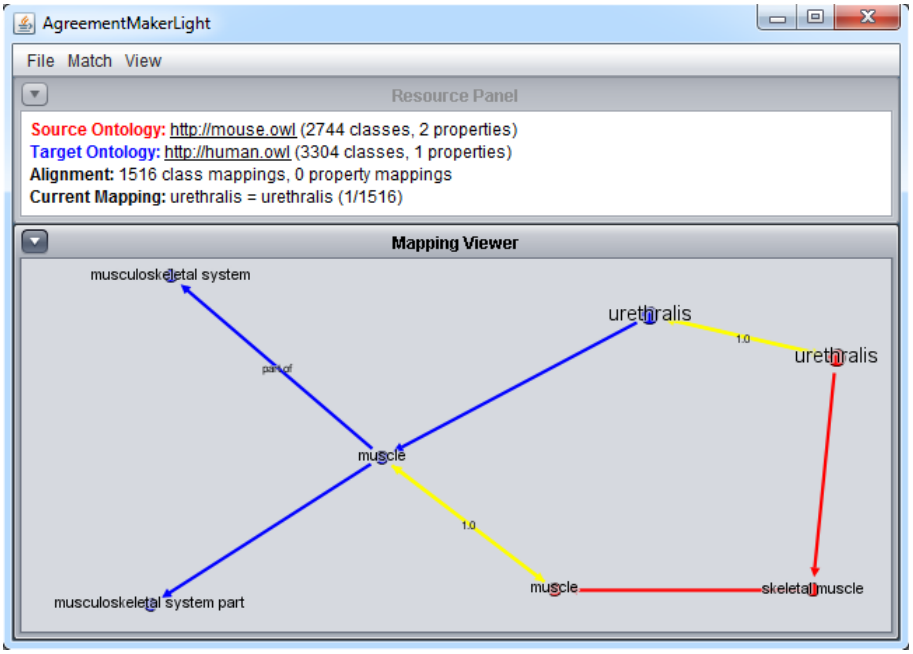
The AML ontology matching framework, which is represented in Figure 1, was designed with scalability in mind. It includes several innovative features to maximize the effectiveness of the matching process while maintaining a reduced complexity, and is divided in three main modules: ontology loading, ontology matching, and alignment selection and repair.

The ontology loading module is responsible for reading ontologies and parsing their information into the AML ontology data structures, which were conceived to enable linear-complexity matching [4]. The most important structure for matching is the *Lexicon*, a table of class names and synonyms in an ontology, which uses a ranking system to weight them and score their matches [7].

The ontology matching module contains AML's ontology matching algorithms, or matchers. These are divided into two primary and secondary matchers, with the former being linear-complexity matchers that be employed globally in all matching problems and the latter being polynomial-complexity matchers that can only be applied locally on large problems. The use of background knowledge in primary matchers is a key feature in AML, and it includes a novel automated background knowledge selection algorithm.

The alignment selection and repair module ensures that the final alignment has

[1] https://github.com/AgreementMakerLight



Source Ontology: http://mouse.owl (2744 classes, 2 properties)
Target Ontology: http://human.owl (3304 classes, 1 properties)
Alignment: 1516 class mappings, 0 property mappings
Current Mapping: urethralis = urethralis (1/1516)

Fig. 2. AgreementMakerLight graphical user interface.

the desired cardinality and that it is coherent (i.e., does not lead to the violation of restrictions of the ontologies) which is important for several applications. AML's approximate alignment repair algorithm features a novel modularization step which identifies the minimal set of classes that need to be analyzed for coherence, thus greatly reducing the scale of the repair problem [8].

2.2 User Interface

The GUI was a recent addition to AML, as we sought to make our system available to a wider range of users. The main challenge in designing the GUI was finding a way to visualize an alignment between ontologies that was both scalable and useful for the user. Our solution was to visualize only the neighborhood of one mapping at a time, while providing several options for navigating through the alignment [6]. The result is a simple and easy to use GUI which is shown in Figure 2.

3 AgreementMakerLight in Use

Despite only being in development since 2013, AML already achieved top results in that years' edition of the Ontology Alignment Evaluation Initiative (OAEI) [3]. Namely, it ranked first in F-measure in the anatomy track, and second in the

large biomedical ontologies track and also in the conference track. In addition to its effectiveness in matching life sciences ontologies, AML was characterized by a high F-measure/run time ratio, which attests to its efficiency.
AML has also been used to analyze ontology reference alignments [5], and is currently being used to match chemical and pharmaceutical ontologies, and to analyze the coherence of BioPortal mappings.
AML is easy to use, thanks to its GUI, and also very efficient. In a current personal computer [1] AML's matching procedure takes from under 1 minute for medium-sized problems (up to 10,000 classes per ontology) to at most 20 minutes for very large matching problems (up to 100,000 classes per ontology).

Acknowledgments

DF, CP, ES and FMC were funded by the Portuguese FCT through the SOMER project (PTDC/EIA-EIA/119119/2010) and the LASIGE Strategic Project (PEst-OE/EEI/UI0408/2014). The research of IFC was partially supported by NSF Awards CCF–1331800, IIS–1213013, IIS–1143926, and IIS–0812258 and by a UIC-IPCE Civic Engagement Research Fund Award.

References

1. I. F. Cruz, F. Palandri Antonelli, and C. Stroe. AgreementMaker: Efficient Matching for Large Real-World Schemas and Ontologies. *PVLDB*, 2(2):1586–1589, 2009.
2. J. Euzenat and P. Shvaiko. *Ontology Matching*. Springer-Verlag New York Inc, 2007.
3. D. Faria, C. Pesquita, E. Santos, I. F. Cruz, and F. M. Couto. AgreementMakerLight Results for OAEI 2013. In *ISWC International Workshop on Ontology Matching (OM)*, 2013.
4. D. Faria, C. Pesquita, E. Santos, M. Palmonari, I. F. Cruz, and F. M. Couto. The AgreementMakerLight Ontology Matching System. In *OTM Conferences*, volume 8185 of *LNCS*, pages 527–541, 2013.
5. C. Pesquita, D. Faria, E. Santos, and F. M. Couto. To repair or not to repair: reconciling correctness and coherence in ontology reference alignments. In *ISWC International Workshop on Ontology Matching (OM)*, 2013.
6. C. Pesquita, D. Faria, E. Santos, and F. M. Couto. Towards visualizing the alignment of large biomedical ontologies. In *10th International Conference on Data Integration in the Life Sciences*, 2014.
7. C. Pesquita, D. Faria, C. Stroe, E. Santos, I. F. Cruz, and F. M. Couto. What's in a 'nym'? Synonyms in Biomedical Ontology Matching. In *The Semantic Web - ISWC 2013*, volume 8218 of *Lecture Notes in Computer Science*, pages 526–541. Springer Berlin Heidelberg, 2013.
8. E. Santos, D. Faria, C. Pesquita, and F. M. Couto. Ontology alignment repair through modularization and confidence-based heuristics. *CoRR*, arXiv:1307.5322, 2013.

[1] With Pentium® Core™ i5 or i7 processor or equivalent, and at least 8GB RAM

Towards a Linked Biology – An integrated perspective of phenotypes and phylogenetic trees

Eduardo Miranda[1], Anaïs Grand[2], Régine Vignes Lebbe[3], and André Santanchè[1]

[1] Institute of Computing – State University of Campinas
Av. Albert Einstein, 1251 – Cidade Universitária, Campinas, Brazil
[2] Muséum national d'histoire naturelle, CR2P - UMR 7207 CNRS/MNHN/Univ Paris 06, 57 rue Cuvier, CP48 - F-75005, Paris, France
[3] UPMC Univ Paris 06, CR2P - UMR 7207 CNRS/MNHN/Univ Paris 06, 4 Place Jussieu, Tour 46-56, 5ème étage, F-75005, Paris, France
eduardo.dpm@gmail.com,grandanais@gmail.com
regine.vignes_lebbe@upmc.fr,santanche@ic.unicamp.br

Abstract. A large number of studies in biology, including those involving phylogenetic tree reconstruction, result in the production of a huge amount of data – e.g., phenotype descriptions, morphological data matrices, etc. Biologists increasingly face a challenge and opportunity of effectively discovering useful knowledge by crossing and comparing several pieces of information, not always linked and integrated. Our motivation stems from the idea of transforming these data into a network of relationships, looking for links among related elements and enhancing the ability to solve more complex problems supported by machines. This work addresses this problem through a graph database model, linking and coupling phylogenetic trees and phenotype descriptions. In this paper we give an overview of an experiment exploiting the synergy of linked data sources to support biologists in data analysis, comparison and inferences.

1 Introduction

In spite of several initiatives to publish open data and to combine phylogenetic trees, there is still a high amount of latent knowledge hidden in potentially linkable data, which are fragmented in several heterogeneous datasources. This heterogeneous multitude of resources can be seen as a dataspace [1], where pieces of data maintain unexploited potential links. This work addresses this problem in a specific scenario. We gather together in a graph database data coming from distinct sources, containing phenotype descriptions and phylogenetic trees. This graph subsidizes links discovery, aimed at supporting biologists in the analysis and comparison of phylogenetic information (such as homology hypotheses, characters and trees) of hypothetical phylogenetic trees.

This paper is organized as follows: Section 2 presents our three layer method and the architecture of our system; Section 3 presents our graph-based model, illustrates our approach to discover links based on similarity and the results of the visualization tool; Section 4 presents concluding remarks.

2 Three Layer Method and System Architecture

In this research we propose a Three Layer Method, in which fragmentary data sources are mapped to a graph database, where the data will be pass through integration steps targeting an ontology. Our approach remodels sources from the dataspace to a graph representation, in which the data can be unified and linked, subsidizing the discovery of latent knowledge, which raises from the relations. The graph model was designed to be published on the Web in a Linked Data approach. Graph transformations will be applied for the transition from representations in the dataspace to a more formalized representation through ontologies. This work focuses in the graph representation and its application to support an analytical tool to compare data across studies.

Figure 1 summarizes the general architecture of our system. From a set of heterogeneous data sources (1), we ingest and transform data in a graph (2) stored in a graph database (3). In this stage of the research, we are interested in phenotype descriptions and phylogenetic trees, even though the architecture was designed to afford smooth future extensions to other kinds of biological data. In step (3) each data source will result in a distinct graph. We applied LSIDs to unify Operational Taxonomic Units (OTUs) in the graph referring to the same real world object (4). In step (4), we are developing algorithms to discover relations and find similarities among nodes in the graph, which are made explicit by adding new edges in the graph. The resulting graph can be locally analyzed by a researcher; can be published on the Web in a Linked Data approach to be remotely exploited (6); and will subsidize the expansion and enrichment of ontologies in the future (7).

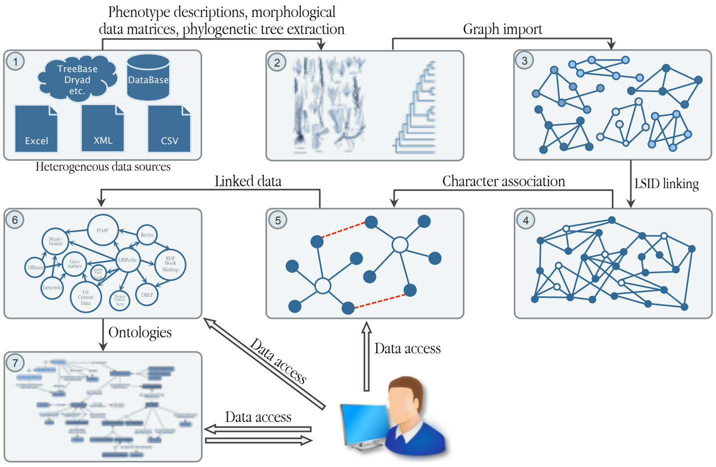


Fig. 1. General System Architecture.

3 Unified Graph Data Model and Link Discovery

In this section we will present an overview of our proposed graph model. From the numerous graph data models proposed, the *property graph* model was adopted in the present work. In *property graph*, nodes and relationships can maintain extra metadata as a set of key/value pairs. Moreover, relationships are typed, enabling to create multi-relational networks with heterogeneous sets of edges.

Figure 2 shows our graph data model. The tables below the nodes/edges represent their types and metadata. Our model maps parts of the Structured Descriptive Data [4] (SDD) format to a graph data model. The SDD format is a XML-based standard for recording and exchanging descriptions of biological and biodiversity data of any type [2]. We mapped as follows: OTUs are entities made nodes. SDD States were mapped to nodes, since equivalent *States* related to several OTUs can be unified and related. Finally, the SDD *StateDefinition* makes a semantic bridge – relationship – between OTUs and characters. This part of the model, is a common denominator of phenotype descriptions, conceived in our previous work [3].

Our model comprises into a single place phenotype description and phylogenetic tree. For this reason a new node called HTU (Hypothetical Taxonomic Unit) is present in this model. HTUs are internal nodes in phylogenetic trees that represent an inferred ancestral organism.

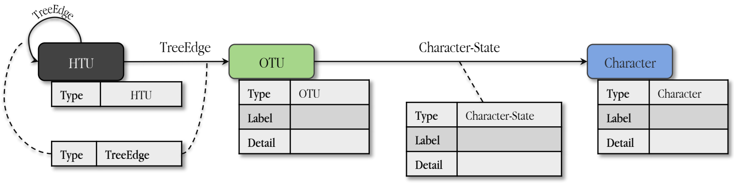


Fig. 2. Property Graph Model

In order to illustrate the possibilities raised by the unification and linking of data of phenotype descriptions with phylogenetic trees, we developed a practical experiment executed in our system, which involves the linking discovery among characters. In this respect, we are proposing a heuristic similarity measure that computes the similarity degree between two morphological character. The similarity degree is based on a weighted combination of factors concerning two characters: their labels; whether they describe the same taxa; the common or similar character-states.

Figure 3 shows the result of a visual tool that creates an edge between each two related characters based on the similarity measure. This is a simple but powerful visualization tool that could play a pivotal role in supporting biologists to understand and detect correlation between characters. That tool was able to show a graph clique among highly related characters – see figure 3. Knowing that characters are similar to some extent can encourage biologists to unify identical relationships.

[4] The Structured Descriptive Data format (*http://wiki.tdwg.org/SDD)*

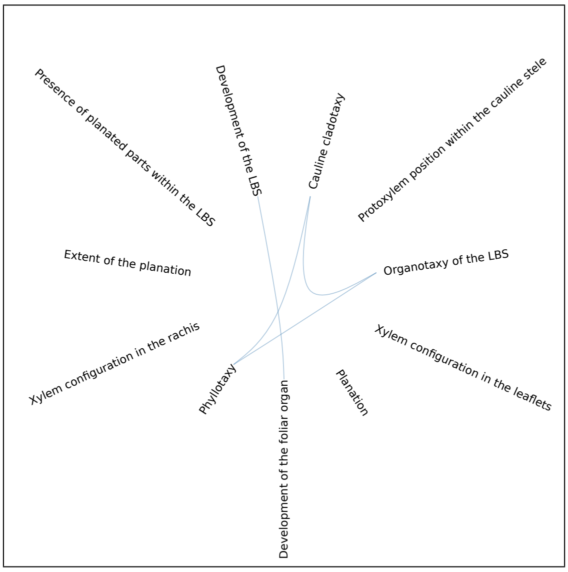

Presence of planated parts within the LBS
Development of the LBS
Cauline cladotaxy
Protoxylem position within the cauline stele
Extent of the planation
Organotaxy of the LBS
Xylem configuration in the rachis
Xylem configuration in the leaflets
Phyllotaxy
Development of the foliar organ
Planation

Fig. 3. Practical Experiment

4 Conclusion

Our unified model enabled us to discover and make explicit the potential semantics raised by linking previously unconnected information. The viability and the potential of our approach were tested by experiments in which 2 distinct authors descriptions of fossils were inserted into a graph database and analyzed by the similarity measure method mentioned in this paper. We developed a tool to visualize how close related are two given characters and some preliminary results are presented. This tool has the potential to indicate the recurring use of the same character in different studies and might support biologists to understand and detect correlation between characters.

Acknowledgments. Work partially financed by (CNPq 138197/2011-3), the Microsoft Research FAPESP Virtual Institute (NavScales project), CNPq (MuZOO Project and PRONEX-FAPESP), INCT in Web Science(CNPq 557.128/2009-9) and CAPES-COFECUB, as well as individual grants from CNPq.

References

1. Franklin, M., Halevy, A., Maier, D.: From databases to dataspaces: a new abstraction for information management. SIGMOD Rec. **34**(4) (December 2005) 27–33
2. Hagedorn, G.: Structuring Descriptive Data of Organisms – Requirement Analysis and Information Models. PhD thesis, Universität Bayreuth,Fakultät für Biologie, Chemie und Geowissenschaften (11 2007)
3. Miranda, E., Santanchè, A.: Unifying phenotypes to support semantic descriptions. VI Brazilian Conference on Ontological Research (Ontobras) (09 2013)

Developments and Perspectives for CABRI Web Services

Daniele Pierpaolo Colobraro, Paolo Romano
IRCCS AOU San Martino IST, Genova, Italy
{danielepierpaolo.colobraro,paolo.romano}@hsanmartino.it

Abstract. In this short paper, we introduce the on-going improvements of CABRI Web Services. These are meant to make programmatic access to information on microbial resources much simpler and more effective, by also enabling a real interoperability with some of the most used bioinformatics databases. Such improvements mainly consist in an extension of the returned information, the adoption of the Microbial Common Language (MCL) for data exchange, and of the implementation of REST standard along with the already available SOAP standard.

Keywords: microbial biological resources; web services; data interoperability; data integration.

1 Introduction

Information provided by Biological Resource Centers (BRCs) is of increasing interest for researchers. Such information consists in catalogues of well conserved and characterized micro-organisms that can be requested from BRCs. These catalogues are made available through BRC web sites, but some portals enables an integrated access to many catalogues, e.g. the Common Access to Biological Resources and Information (CABRI) [1] (http://www.cabri.org/) and the Global Catalogue of Microorganisms (GCM) [2] (http://gcm.wfcc.info/). CABRI network services offer access to 28 catalogues from European BRCs, including more than 120,000 resources. They were implemented in 1999 as a final deliverable of the CABRI EU project. Catalogues are submitted in a standard format for implementation in a common SRS site.

Much more information exist on these micro-organisms and both CABRI and GCM offer links to external resources, including PubMed, for literature, and the European Nucleotide Archive (ENA), for sequences. Other systems, such as StrainInfo [3] (http://www.straininfo.net/), can offer useful information about strains.

The Microbial Resource Research Infrastructure (MIRRI) is a recently funded European infrastructure, still in its preparatory phase [4] (http://www.mirri.org/). One of its workpackages aims at defining methods for improving the quantity and quality of information on microbial resources, designing a new portal for the distribution of collections' catalogues, and assessing best ways to improve interoperability with other databases, so that improved software in critical domains, like health and environment, may be developed by leveraging on the future MIRRI information system.

Indeed, a new generation of software able to improve interoperability of biomedical information systems would be useful to support advanced research. Software technologies like Web Services and Workflow Management Systems are being increasingly adopted [5,6,7,8]. To this aim, CABRI Web Services (CABRI-WS) [9] (http://bioinformatics.istge.it/ibws/doc_cabri.html) and the Microbiological Common Language (MCL), an XML based data exchange format for microbiological information [10] (http://www.straininfo.net/projects/mcl) may be essential components for biological resources. CABRI-WS are included in the IST Bioinformatics Web Services (IBWS) that were deployed at the National Cancer Research Institute of Genoa, now IRCCS AOU San Martino IST. IBWS also include SWS (SRS by Web Services), that allow to query databases included in SRS sites, and TP53 Web Services, that retrieve data from an SRS implementation of the IARC TP53 Mutation Database.

Here, we introduce some of the on-going improvements of CABRI Web Services.

2 Methods

Distinct WS are available for each biological resource type in CABRI: bacteria and archaea strains, filamentous fungi strains, yeasts strains, plasmids, phages, and human and animal cell lines. Resources can be searched by name, identifier or using free text. Two types of services were implemented: i) searching for name or free text and returning IDs, and ii) searching for an ID and returning full records. The output consists in the contents of CABRI catalogues, formatted as flat files (see Box 1).

```
Strain_number LMG 1(t1)
Other_collection_numbers CCUG 34964; NCIB 12128
Restrictions Biohazard group 1
Organism_type Bacteria
Name Phyllobacterium myrsinacearum, (ex Knösel 1962) Knösel 1984VL
  emend. Mergaert, Cnockaert and Swings 2002 VP
Infrasubspecific_names -
Status -
History <- 1973, D.Knösel (Phyllobacterium rubiacearum)
Conditions_for_growth Medium 1, 25C
Form_of_supply Dried
Isolated_from Pavetta zimmermannia
Geographic_origin Germany, Stuttgart-Hohenheim
Remarks Stable colony type isolated from LMG 1. Type strain of
  Phyllobacterium rubiacearum. See also Agrobacterium sp. LMG 1(t2)
```

Box 1. CABRI original flat file format for LMG 1(t1)

CABRI-WS were deployed by using Soaplab, a tool able to provide programmatic access to local, command-line applications and to the contents of ordinary web pages, whose only requirements are Apache Tomcat with the Axis toolkit and a Java Virtual Machine. In Soaplab, new Web Services are added to the system by defining simple descriptions of related execution commands. These are written in the AJAX Command Definition (ACD) language. CABRI WS can be accessed through any WSDL-SOAP compliant software, including the Taverna Workbench.

CABRI-WS improvements are aimed at: i) extending returned information, ii) adopting MCL for data exchange, iii) implementing REST based Web Services.

We are extending information returned by CABRI-WS by adding data that is available in CCINFO, a directory of culture collections provided by the World Data Center for Microorganism (WDCM), Straininfo, and some reference information systems. MCL is able to represent the contents of CABRI catalogues, with a greater precision and it may be used for the output of the new CABRI-WS. See examples in Box 2. Due to the extension of CABRI-WS data, a revision of MCL may be needed.

(a)

```
<mcl:BCR>
   <mcl:WDCMNumber>296</mcl:WDCMNumber>
   <mcl:fullName>Belgian Coordinated Collections of Microorganisms/
LMG Bacteria Colletion</mcl:fullName>
   <mcl:acronym>LMG or BCCM/LMG</mcl:acronym>
</mcl:BCR>
```

(b)

```
<mcl:Culture>
   <mcl:strainNumber>LMG 1(t1)</mcl:strainNumber>
   <!-- Strain number from Cabri -->
   <mcl:otherStrainNumber>CCUG 34964</mcl:otherStrainNumber>
   <!-- Strain number from Cabri -->
   <mcl:otherStrainNumber>NCIB 12128</mcl:otherStrainNumber>
   <!-- Strain number from straininfo.net -->
   <mcl:otherStrainNumber>CECT 4452</mcl:otherStrainNumber>
   <!-- Strain number from straininfo.net -->
```

Box 2. Information on the BRC from CCINFO (a) and extended information on other strain numbers from Strainfo (b), both excerted from the oputput fomatted with MCL.

Web Service	**Output**	**Call** (prefix http://www.cabri.org/cws/)
getCataloguesList	List of catalogues	getCataloguesList/views
getCatalogue	List of all strains in a given catalogue	getCatalogue/views? name=CABI_BACT
getData	Query catalogues	getData/views?resname=bacillus subtilis

Table 1. Summary of REST Web Services

The REST standard has proven to be both effective and simple to adopt. Improved CABRI-WS are being implemented through this standard. To this aim, the WS will be implemented through an Apache web server. Scripts are presently being written in python. In Table 1, a summary of available WS is shown. More information is being made available on-line at http://bioinformatics.istge.it/ibws/rest.html.

3 Conclusion

We have presented the main improvements of CABRI WS which are being carried out in the context of the MIRRI project. The new WS include an extended contents, by incorporating data from other information systems. A REST interface is being built and will flank the current SOAP one. The search and retrieval approaches reproduce the features of the standard web interface. The output is based on the MCL language for representation and exchange of microbiological information. These improvements are aimed at simplifying and making interoperability of microbial information more effective, which is one of the main aims of the MIRRI project.

Acknowledgements. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 312251.

References

1. Romano, P., Kracht, M., Manniello, M.A., Stegehuis, G., Fritze, D.: The role of informatics in the coordinated management of biological resources collections. Appl Bioinf. 4(3):175-186, 2005.
2. Wu L, Sun Q, Sugawara H, Yang S, Zhou Y, McCluskey K, Vasilenko A, Suzuki K-I, Ohkuma M, Lee Y, Robert V, Ingsriswang S, Guissart F, Desmeth P, Ma J. Global catalogue of microorganisms (gcm): a comprehensive database and information retrieval, analysis, and visualization system for microbial resources. BMC Genomics 2013, 14:933.
3. Dawyndt P, Vancanneyt M, De Meyer H, Swings J. Knowledge accumulation and resolution of data inconsistencies during the integration of microbial information sources, IEEE Transactions on Knowledge and Data Engineering, vol.17, no.8, pp. 1111-1126, Aug. 2005
4. Schüngel M, Stackebrandt E, Bizet C, Smith D. MIRRI The Microbial Resource Research Infrastructure: managing resources for the bio-economy. EMBnet.journal 2013, 19(1):5-8.
5. Romano P. Automation of in-silico data analysis processes through workflow management systems, Briefings in Bioinformatics 2008 9(1):57-68.
6. Wolstencroft K, Haines R, Fellows D, Williams A, Withers D, Owen S, Soiland-Reyes S, Dunlop I, Nenadic A, Fisher P, Bhagat J, Belhajjame K, Bacall F, Hardisty A, Nieva de la Hidalga A, Balcazar Vargas MP, Sufi S, Goble C. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. Nucleic Acids Research 2013, 41(W1): W557-W561.
7. Bhagat J, Tanoh F, Nzuobontane E, Laurent T, Orlowski J, Roos M, Wolstencroft K, Aleksejevs S, Stevens R, Pettifer S, Lopez R, Goble C. BioCatalogue: a universal catalogue of web services for the life sciences. Nucl. Acids Res. 2010, 38: 689-694.
8. Romano P, Marra D, Milanesi L. Web services and workflow management for biological resources, BMC Bioinformatics 2005, 6(Suppl 4):S24.
9. Zappa A, Miele M, Romano P. IBWS: IST Bioinformatics Web Services. Nucleic Acids Research 2010, 38(Web Server issue):W712-W718.
10. Verslyppe B, Kottman R, De Vos P, De Baets B, Dawyndt P. Microbiological Common Language (MCL): a standard for electronic information exchange in the Microbial Commons. Research in Microbiology 2010, 161(6), 439-445.

Towards a Plant Experimental Assay Ontology

Nuno D. Mendes[1,2], Pedro T. Monteiro[1], Cátia Vaz[1,3], and Inês Chaves[1,4]

[1] INESC-ID - Instituto de Engenharia de Sistemas e Computadores, Lisboa, PT
{ndm,ptgm}@kdbio.inesc-id.pt,
[2] IBET - Instituto de Biologia Experimental Tecnolégica, Oeiras, PT
[3] ISEL - Instituto Superior de Engenharia de Lisboa, Lisboa, PT
cvaz@cc.isel.ipl.pt,
[4] ITQB - Instituto de Tecnologia Química e Biológica, Oeiras, PT
ichaves@itqb.unl.pt

Abstract. The Plant domain has been the subject of several attempts to structure and formally define terms and corresponding relations, such as their anatomical features, developmental stages, and the application of particular experimental procedures to a biological problem. However, a focus on experimental assays in order to describe the whole experimental procedure, to the best of our knowledge, has only been attempted in the context of a very general description based on classical views of the scientific method [1]. In this study, we focus on the development and proposal of an ontology dedicated to the description of these experimental procedures, regardless of the scientific questions that prompted the assays. This ontology includes entities from three distinct realms (biological, physical and data), which include both experimental products, their relations and the protocols describing their manipulation. The final outcome is a useful and comprehensive ontology in the plant domain, to be used as a log book by experimentalists, providing a formal relation between entities.

Keywords: Ontology, Modeling, Biology, Experimental assays

Introduction

Ontologies originated from the need to formally specify a controlled set of terms and their relationships in the context of a knowledge domain. The advantages of this type of approach include the ability to share structured information between different users and software tools, to reuse the established vocabulary, and, not less importantly, to make domain assumptions explicit [2].

Following their inception as a formal knowledge representation technique, several tools have since surfaced which elicit their practical use, including sophisticated ontology editors [3], semantic databases and query languages.

The wealth of data generated by contemporary biological experimental studies presents several challenges, not only due to its magnitude, but also to its heterogeneity and interdependence. Additionally, the diversity of protocols, tools and data formats, as well as the several different context-specific parameters used at different steps render the fundamental requirement of experimental reproducibility much more difficult to accomplish. This calls for an attempt to meaningfully represent the experimental procedures as well as the data they produce.

Here we focus our attention on experimental assays designed for studying the plant domain. Studies with other types of subjects (e.g animals or bacteria) may be significantly different, which justifies our choice to restrict the scope of our efforts.

Problem and Designed Solution

Several ontologies describing developmental and anatomical characteristics of plants, their environment and even the types of stress they can be subject to have already been proposed (e.g. Plant Ontology (PO) [4], Plant Trait Ontology (TO) [5], Plant Infectious Diseases (IDOPlant) [6]). Although ontologies specifically dedicated to the description of experimental procedures in general do exist [1], their foremost concern is the description of experimental design, hypothesis testing and the ultimate goal of the experiments. The ontology proposed here, on the other hand, is mainly dedicated to the description of the pipeline of manipulations performed from specimens to data.

This focus on the actual experimental procedure rather than the teleological and epistemological foundations of the assay is justified by the observation that the data produced by experimental assays may be, and is, generally used in several subsequent studies with, presumably, different objectives and that there is much data produced outside the scope of a clear experimental design or a particular biological question (e.g. genome sequencing). There are, however, several opportunities to reuse preexisting ontologies, namely in the description of the growth environment [7], designation of the fractioned samples [4], or the modelling of the notion of time when referring to operations in general [8].

Our ontology, which is currently in active development (summarised in Figure 1), considers essentially three types of entities: biological entities, physical entities and data entities. An additional term - material entity - is used to group two top-level concepts from the biological and physical realms, for convenience. An ancillary ontology was also created to describe iterative processing pipelines (see Figure 2).

Biological entities refer to biological material or to manipulations thereof. A central concept is that of Biological Subject, which can be a specimen or a specimen pool (a single plant, or a group of plants, which can be several individual clones, individuals of different species, etc), or a Biological Sample, which is a transformation of an original Biological Subject. These transformations can be either experimental manipulations (e.g. biotic or abiotic stress) or the result of fractioning (e.g. the isolation of a particular tissue or another anatomical feature). Successive Biological Samples can be obtained by applying a sequence of transformations. Physical entities refer to non-living material, usually the result of an extraction procedure (e.g. nucleic acids extraction, protein extraction, etc), or the product of an analysis protocol (e.g. a gel obtained from electrophoresis), or the manipulations thereof. Data entities, likewise, refer to informational concepts and their manipulation. Here Data Subjects are organized in subclasses related to a particular *DataAcquisitionProtocol*. Thus, *ImageData* is obtained using an *ImageAcquisitionProtocol*, *ProteinData* with, for instance, *MassSpectrometry*, and so forth.

Some of the above mentioned entities have object properties characterizing their relationships and effectively imposing restrictions between the instances of each entity. Examples of these properties are: a *PhysicalAggregate* which contains one or more *PhysicalSubject*, or a *Specimen* which has a *Taxonomy* and it contains one or more *SpecimenPool*.

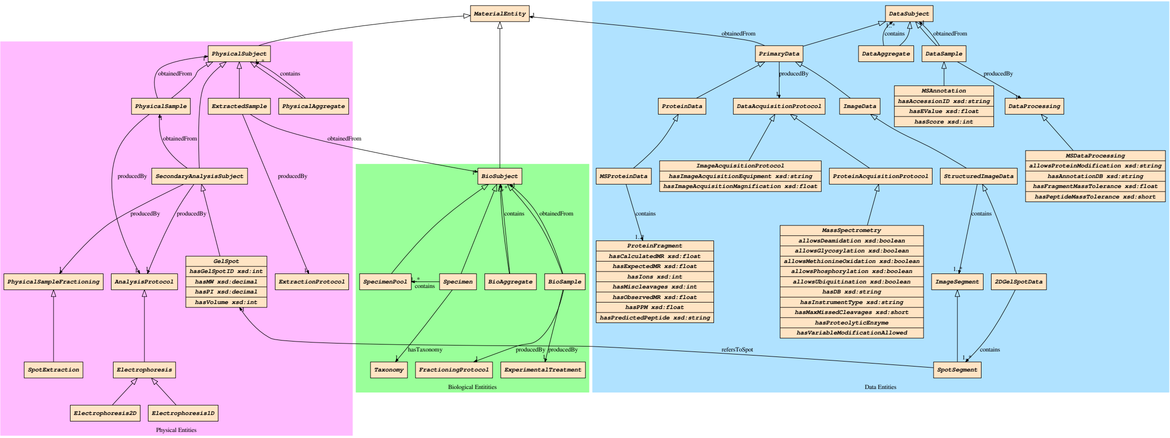


Fig. 1. Ontology representing experimental products, their relations and protocols divided into three main categories of entities: biological, physical and data. A full description of the proposed ontology (generated by Protégé's OWLDoc plugin) can be found at http://kdbio.inesc-id.pt/PEAO/.

The distinctions we here present about the nature of these entities (biological, physical or data) are not explicitly distinguished by the ontology, but their interdependence is constrained by the domain and range of the object properties it defines.

Ongoing Work and Ontology Implementation

As an example illustrating the usefulness of our proposition, we considered the data collected from the interactions between coffee and coffee leaf rust. The data used as test-case is the description of the experimental design and stress imposition, the cytological results, protein gel electrophoresis images, mass spectrometry results and protein assignment results with annotation and extraction and analysis protocols. These data is continually being produced by the Centro de Investigação das Ferrugens do Cafeeiro of Instituto de Investigação Científica Tropical (IICT), and it has been analysed by Instituto

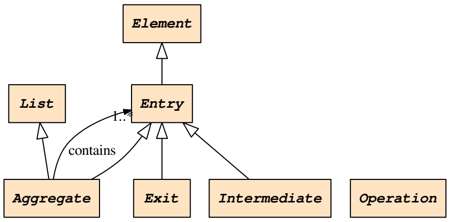


Fig. 2. Auxiliary ontology describing iterative process pipelines.

de Tecnologia Qumica e Biológica/Instituto de Biologia Experimental Tecnológica (iBET/ITQB) in collaboration with Instituto de Engenharia de Sistemas e Computadores (INESC-ID).

The plant experimental assay ontology and the pipeline patterns ontology, are being developed using the modelling development tool Protégé [3]. This tool is widely used for modelling a wide range of domains, providing tools for visualization, ontology merging, refactoring or embedded reasoners for the logical verification of the ontology.

The parameters of the several types of manipulations of biological, physical and data entities are given as data properties of those manipulations. For instance, a *SpotSegment*, a special case of *Data Entity*, refers to a *GelSpot*, a special case of *Physical Entity*. The latter contains four data properties: *hasGelSpotID*, *hasMW*, *hasPI* and *hasVolume*, describing the results of an electrophoresis. Each electrophoresis, which is a particular type of an analysis protocol that can be performed to a *Physical Entity*, will then contain multiple *GelSpot*. On the other hand, a *SpotSegment* is a special case of an *ImageSegment* which is part of a *StructuredImage*, corresponding to an image of the actual gel produced by electrophoresis.

Conclusions

This work is a first attempt at enabling the creation of a repository of data produced by plant experimental assays. The representation of the data using an ontology elicits the preservation of the semantic relationships between the entities represented therein, which facilitates the interpretation of the results and the integration of data produced by different experiments. Additionally, the ontology is easily extensible to incorporate new types of data and experimental manipulations and can, therefore, accommodate a growing number of different experimental procedures.

Acknowledgments: This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under projects PEst-OE/EEI/LA0021/2013 and DataStorm EXCL/EEI-ESS/0257/2012.

References

1. L.N. Soldatova, R.D. King: An Ontology of Scientific Experiments. J. R. Soc. Interface. 3(11), 795–803 (2006)
2. Michael Hammer, Dennis McLeod: Database Description with SDM: A Semantic Database Model. TODS 6(3), 351–386 (1981)
3. J.H. Gennari, M.A. Musen, R.W. Fergerson, W.E. Grosso, M. Crubezy, H. Eriksson, N.F. Noy, S.W. Tu: The Evolution of Protégé: an Environment for Knowledge-based Systems Development. Int. J. Hum. Comput. Stud. 58, 89–123 (2003)
4. Ilic K., E. A. Kellogg, P. Jaiswal, F. Zapata, P. F. Stevens, L. Vincent, S. Avraham, *et al.*: The Plant Structure Ontology, a Unified Vocabulary of Anatomy and Morphology of a Flowering Plant. Plant Physiology 143:587–599 (2007)
5. Jaiswal P.: Gramene Database: A Hub for Comparative Plant Genomics. Methods Mol. Biol. 678:247–275 (2011)
6. Walls R. L., B. Smith, J. Elser, A. Goldfain, D. W. Stevenson, and P. Jaiswal: A Plant Disease Extension of the Infectious Disease Ontology. Proc. 3rd Intl. Conf. Biomedical Ontology. (2012)
7. http://www.ontobee.org/browser/index.php?o=EO. Accessed January 28, 2014
8. http://www.w3.org/TR/owl-time/. Accessed May 29, 2014

Identifying, Interpreting, and Communicating Changes in XML-encoded Models of Biological Systems

Martin Scharm[1,⋆], Olaf Wolkenhauer[1,2], and Dagmar Waltemath[1]

[1] Department of Systems Biology and Bioinformatics, University of Rostock, Rostock, Germany
[2] Stellenbosch Institute for Advanced Study (STIAS), Wallenberg Research Centre at Stellenbosch University, Stellenbosch 7600, South Africa

Background

Research in systems biology enhanced our knowledge of biological environments. Many discoveries are recorded in computational models which encode the structure of biological networks, and describe their temporal and spatial behavior. Due to tremendous efforts by the research community, the number of openly available models is numerous and still continually increasing [1]. To support the sharing of models and, thus, the reuse of research results, repositories such as the BioModels Database [2] and the Cellml Model Repository [3] collect and store models in exchangeable formats such as the Systems Biology Markup Language (SBML, [4]), or CellML [5]. Since only accessible models can be reused, such repositories are essential to guarantee transparent research.

However, model repositories to date lack sufficient mechanisms to track the updates of models in their databases [6]. Model versions often cannot be addressed unambiguously and changes occurring between versions of a model are not communicated transparently. Therefore, a framework to identify the differences between models and their versions is a fundamental requirement to compare and combine models. Only with difference detection at hand users are able to grasp a model's history and to identify errors and inconsistencies.

Results

On the poster, we reflect on the following requirements for systems that provide version control for models:

- All versions of a model must be accessible.
- Information must be available on when a model changed, how, why, and by whom.
- Changes in model versions must be made transparent to the user.

[⋆] To whom correspondence should be addressed

Our current research concentrates on developing efficient and reliable difference detection for versions of models. We thereby address the abovementioned requirement that information must be available on *how* a model changed over time. Specificially, our algorithm for difference detection, BiVeS[3], is applicable to models encoded in SBML or CellML. As standard representation formats for computational models in biology use XML, BiVeS bases on an XML-diff algorithm, namely the XyDiff algorithm [7]. BiVeS identifies structures in the XML trees that both documents have in common and maps their subgraphs onto each other. The resulting mapping is then propagated into the rest of the tree, possibly leading to further mappings. That way, moved entities can be identified, as well as inserts and deletes. The algorithm is furthermore format-specific in the sense that it respects the structure of the representation formats. The major elements of the SBML Level 3 specification [8], for example, are biological entitities (species) that participate in biological processes (reactions). CellML very generally encodes biological facts as sets of interacting components. Both representation formats use semantic annotations (i. e., links to ontologies) to further describe the biological meaning of the single XML elements [9]. We use this information to further improve the mappings. The final set of of differences can be exported in both machine and human readable formats: BiVeS produces an XML-encoded patch containing all modifications which occurred between the two versions of a document (see Figure 1). Changes between model versions are also summarized in a report and highlighted in a graph, which comprehensively displays the updates affecting the reaction network. The algorithm is implemented in a Java library.

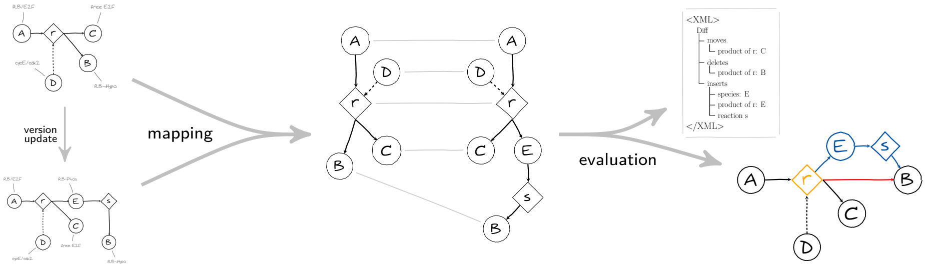
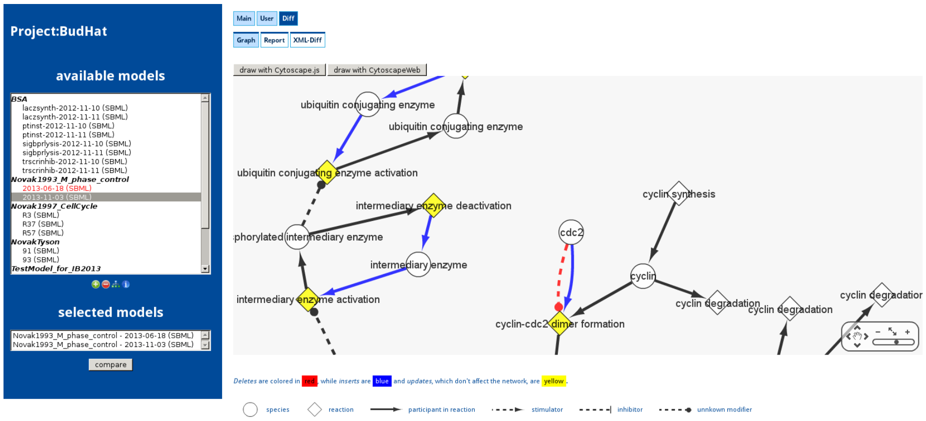


Fig. 1. A sketch of the BiVeS algorithm for difference detection.

Gaining insights into the process of development of a particular model has the potential to increase the confidence in this model and supports the collaboration of distinct research projects dramatically. Consequently, existing model repositories can benefit from extending their software and functionalities with version control. On our poster, we show how the BiVeS library can be integrated

[3] Biomodel Version Control System, `https://sems.uni-rostock.de/projects/bives/`

with existing software: (i) BiVeS offers an API that can be used from other Java tools, (ii) a web service provides access to BiVeS via HTTP, (iii) the library can be executed directly from the command line. BiVeS is already implemented in the Functional Curation project of Chaste [10]. Furthermore, we are currently in touch with the maintainers of SEEK, a data management platform for the life sciences [11], BioModels Database, and the CellML model repository to integrate BiVeS into their infrastructures. On the poster, we demonstrate BiVeS' capabilities with our prototypic web based user interface BudHat[4]. BudHat uses BiVeS to detect changes between versions of a model stored in a database backend. Identified differences are processed and presented human readably. Changes in reaction networks, for example, are highlighted in different colors. An example is shown in Figure 2.



BSA
laczsynth-2012-11-10 (SBML)
laczsynth-2012-11-11 (SBML)
ptinst-2012-11-10 (SBML)
ptinst-2012-11-11 (SBML)
sigbprlysis-2012-11-10 (SBML)
sigbprlysis-2012-11-11 (SBML)
trscrinhib-2012-11-10 (SBML)
trscrinhib-2012-11-11 (SBML)
Novak1993_M_phase_control
2013-06-18 (SBML)
2013-11-03 (SBML)
Novak1997_CellCycle
R3 (SBML)
R37 (SBML)
R57 (SBML)
NovakTyson
91 (SBML)
93 (SBML)
TestModel_for_IB2013

Novak1993_M_phase_control - 2013-06-18 (SBML)
Novak1993_M_phase_control - 2013-11-03 (SBML)

Deletes are colored in red, while *inserts* are blue and *updates*, which don't affect the network, are yellow.

Fig. 2. Screenshot of our prototype BudHat. BudHat is an online tool that displays the differences between model versions, as computed by BiVeS, in some human readable formats (here: highlighted reaction network).

Finally, we discuss first statistics about the evolution of computational models in open repositories. We analysed models from the BioModels Database (144,253 models in SBML format) and the CellML Model Repository (600 different exposures with CellML models). Indeed, models in open repositories do change over time predominantly in two ways: First, models are modified if the representation format, used to encode the model, gets updated. These updates affect a large number of models and form a clear pattern in our visualisation of model changes. For example, all models in BioModels Database where updated when SBML replaced its own standard for links to external resources, MIRIAM, by the identifiers.org scheme [12]. Second, published models are continuously improved and corrected by model curators. We observed updates in the links pointing to terms in bio-ontologies, and to the model's network structure. For

[4] `http://budhat.sems.uni-rostock.de`

the Repressilator model[5], for example, we see that the change in network structure actually affected the simulation outcome. We also identified patterns in the CellML Model Repository, and will discuss possible reasons on our poster. With respect to performance, we used the above data sets to compare our own algorithm for difference detection against the standard Unix diff tool. Unix diff to date is the standard method to compare versions of models in open repositories. However, our results confirm that BiVeS indeed outperforms Unix' diff tool and improves the results obtained by standard XML Diff tools.

Summary

In summary, our poster introduces ongoing research in model management for computational biology, with a focus on the advantages of sophisticated model version control. We discuss in detail the requirements, show our latest research results in terms of algorithm design and tool support, and we present first statistics on the types and frequency of changes in models published in open repositories.

References

1. Henkel *et al.*: Ranked retrieval of computational biology models. *BMC bioinformatics*, 2010.
2. Li *et al.*: BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology*, 2010.
3. Lloyd *et al.*: The CellML Model Repository. *Bioinformatics*, 2008.
4. Hucka *et al.*: The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models.. *Bioinformatics* 19.4:524-531, 2003.
5. Cuellar *et al.*: An overview of CellML 1.1, a biological model description language. *Simulation* 79.12, 2003.
6. Waltemath *et al.*: Improving the reuse of computational models through version control. *Bioinformatics* 29.6:742-748, 2013.
7. Cobena *et al.*: Detecting changes in XML documents. 18th International Conference on Data Engineering, 2002.
8. Hucka *et al.*: The systems biology markup language (SBML): language specification for level 3 version 1 Core (Release 1 Candidate). *Nature proceedings*, 2010.
9. Courtot *et al.*: Controlled vocabularies and semantics in systems biology. *Molecular systems biology* 7.1, 2011.
10. Cooper *et al.*: High-throughput functional curation of cellular electrophysiology models. *Progress in Biophysics and Molecular Biology*, 2011.
11. Wolstencroft *et al.*: The SEEK: a platform for sharing data and models in systems biology. *Methods Enzymol* 500:629-655, 2011.
12. Juty *et al.*: Identifiers. org and MIRIAM Registry: community resources to provide persistent identification.. *Nucleic acids research* 40.D1:D580-D586, 2012.

[5] `http://www.ebi.ac.uk/biomodels-main/BIOMD0000000012`