Helena Galhardas
Erhard Rahm (Eds.)

LNBI 8574

Data Integration in the Life Sciences

10th International Conference, DILS 2014
Lisbon, Portugal, July 17–18, 2014
Proceedings

Springer

Lecture Notes in Bioinformatics 8574

Subseries of Lecture Notes in Computer Science

Helena Galhardas Erhard Rahm (Eds.)

Data Integration in the Life Sciences

10th International Conference, DILS 2014
Lisbon, Portugal, July 17-18, 2014
Proceedings

Springer

Volume Editors

Helena Galhardas
Instituto Superior Técnico, University of Lisbon
INESC-ID
Tagus Park
Av. Prof. Dr. Cavaco Silva
2744-016 Porto Salvo, Portugal
E-mail: helena.galhardas@tecnico.ulisboa.pt

Erhard Rahm
Universität Leipzig
Fakultät für Mathematik und Informatik
Institut für Informatik
Augustusplatz 10
04109 Leipzig, Germany
E-mail: rahm@informatik.uni-leipzig.de

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-319-08589-0 e-ISBN 978-3-319-08590-6
DOI 10.1007/978-3-319-08590-6
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014941868

LNCS Sublibrary: SL 8 – Bioinformatics

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume of *Lecture Notes in Bioinformatics* (LNBI) contains the papers presented at DILS 2014: the 10th International Conference on Data Integration in the Life Sciences, held during July 16–17, 2014 in Lisbon, Portugal. In its 10th year, DILS was hosted at Instituto Superior Técnico, University of Lisbon (http://dils2014.inesc-id.pt) and chaired by Erhard Rahm and Helena Galhardas.

The first edition of DILS took place in 2004 in Leipzig, Germany. Over the years, the conference continued to foster discussion, exchange, and innovation in research and development in the areas of data integration and data management in the life sciences. These topics have become more and more important due to the increasing availability of Big Data, coming from high-throughput analytical techniques, large clinical data repositories, biomedical literature and online resources, that offer exciting opportunities and challenges to researchers and professionals from biology, medicine, computer science, and engineering. So far the conference took place in five European countries (Germany, UK, France, Sweden, Portugal), in the USA (three times) and in Canada making DILS a truly international forum.

This year, DILS was a forum that put together invited keynote presentations, oral presentations of peer-reviewed research, application and systems papers, poster and demo presentations. Each submission was reviewed by three Program Committee members. After a careful evaluation process, the Program Committee decided to accept 14 long and short papers that are included in this volume. The accepted papers cover interesting and current topics: data integration platforms and applications, biodiversity data management methods and applications, biomedical ontologies, linked data integration, visualization techniques, and scientific data retrieval and querying. DILS 2014 also included several poster and demo contributions on work-in-progress and system prototypes. The accepted poster and demo papers are published on the conference website.

DILS 2014 featured two distinguished keynote speakers: Dr. Alfonso Valencia and Prof. Jonas S. Almeida. Dr. Alfonso Valencia, vice-director of Basic Research and director of the Structural Biology and Biocomputing Program of the Spanish National Cancer Research Center (CNIO), is an expert in applying computational methods and tools to the analysis of large collections of genomic information, in particular to study protein families and protein interaction networks. His recent research focus is in the domain of cancer (epi)genomics, tumor evolution and precision medicine. In his talk, Dr. Valencia presented the challenges and opportunities of Computational Biology and Big Data. Specifically, he pointed out how technology has influenced the development of Biomedicine and Ecology areas and the current limitations for dealing with large, complex, heterogeneous and low quality data sets and the urge for additional knowledge to interpret the results obtained. Prof. Jonas S. Almeida, Director of the division in Informatics of the Department of Pathology of the University of Alabama at Birmingham (UAB), is specialized on integrative personalized medicine applications. Prof. Almeida has a strong background on all components of quantitative Biology ranging from experimentation, engineering and mathematical modeling to computational statistics and software engineering. His current research interests are on the synergy obtained by combining Semantic Web abstractions and Distributed Cloud Computing approaches to Bioinformatics applications. In his talk, Prof. Almeida overviewed recent solutions in Biomedicine with a particular emphasis on Semantic Web frameworks and code distribution.

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 and additional referees for their excellent contribution in evaluating the submissions. 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 would also 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". We would also like to thank Alfred Hofmann and his team at Springer for their continued cooperation and help in putting this volume together. We also thank the EasyChair team for having developed this tool that enabled us to smoothly manage submissions, reviews and proceedings. 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

Organization

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	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 University, 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

Daniel Faria	Faculty of Sciences, University of Lisbon, Portugal
David Koop	NYU-Poly, USA
Catia Pesquita	Faculty of Sciences, University of Lisbon, Portugal
Emanuel Santos	Faculty of Sciences, University of Lisbon, Portugal
Martin Scharm	University of Rostock, Germany
Cátia Vaz	ISEL, Polytechnic Institute of Lisbon, Portugal

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 University, 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

Keynote Papers

Computational Biology and Big Data: Challenges and Opportunities

Alfonso Valencia

Spanish National Cancer Research Center
valencia@cnio.es

Abstract. Technology is influencing the development of all areas from Biomedicine to Ecology and transforming Biology in a quantitative science. This accelerated technical progression is reflected in the rapid succession of keywords that went in the 20 years from "genomics" to "proteomics", "systems biology" and "synthetic biology" to the current "big data". All of them paving the way to deciphering the function biological systems, from cells to ecosystems, based on the integration of data on genomes, proteomes, metabolomes, environments and conditions.

A promising future that is limited by: a) the current computational technologies for handling large, complex and heterogeneous and in many cases low quality data, and b) very important, the insufficiency of the biological knowledge necessary to interpret the results. In this scenario Bioinformatics and Computational Biology play a central rôle. A particularly good example is the complex task of individual genomes analysis, which involves data organization, integration and interpretation. A challenge that touches many areas of computation and informatics and requires a blend of engineering and scientific developments.

Genome projects are a good example of projects that deal with large scale data, that can be considered part of the Big Data movement. Based on the experience of my group in these projects I will review both the technical framework for handling genomic information and the methods required for the interpretation of the information. In particular, I will focus discuss some of the key scientific problems in the analysis of high-throughput genotype-phenotype information oriented to the prediction of genomics basis of disease conditions.

References

1. Valencia, A., Hidalgo, M.: Getting personalized cancer genome analysis into the clinic: the challenges in Bioinformatics. Genome Medicine, 461 (2012)
2. Vazquez, M., de la Torre, V., Valencia, A.: Cancer Genome Analysis. In: Translational Bioinformatics PLOS Computational Biology Open Access Book, ch. 14 (2012)
3. de Juan, D., Pazos, F., Valencia, A.: Emerging methods in protein co-evolution. Nat. Rev. Genet. 14, 249–261 (2013)
4. Rodriguez, J.M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J.J., Lopez, G., Valencia, A., Tress, M.L.: APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Res. 41, D110–D117 (2013)

5. Ibañez, C., Boullosa, C., Tabarés-Seisdedos, R., Baudot, A., Valencia, A.: Molecular Evidence for the Inverse Comorbidity between Central Nervous System Disorders and Cancers detected by Transcriptomic Meta-analyses. Plos Genet. 10, e1004173 (2014)

The Emergence of the Web Computer: An Hands-on View from the Trenches of Computational Pathology

Jonas S. Almeida

Div. Informatics, Dept. Pathology, Univ. Birmingham at Birmingham
jalmeida@uab.edu

Abstract. The need to contextualize data about an experiment or a patient is increasingly achieved with reference to Big Data resources such as The Cancer Genome Atlas (TCGA). This exercise faces numerous obstacles, from the logistics of traversing a very large and constantly growing data set (the number of files hosted by TCGA doubles every 7 months [1]) to the protection of patient privacy. It also includes an absolute need for "weak AI" to reach domain experts increasingly immersed in mobile platforms. These challenges are not unique to Biomedicine but are, in many regards, particularly difficult to meet in this domain [2]. Correspondingly, the pursuit of solutions is part of the core mission of the new sub-discipline of Computational Pathology [3]. This presentation will overview early stage solutions with applications ranging from image analysis in cytology [4] and sequence analysis [5] to the personalization of cancer treatment. These illustrative applications will be used as part of an argument for the central role played by Web Technologies, with particular emphasis on Semantic Web frameworks and code distribution directly to the ubiquitous Web Platform supported by the modern web browser.

References

1. Robbins, D.E., Grüneberg, A., Deus, H.F., Tanik, M.M., Almeida, J.S.: A self-updating road map of The Cancer Genome Atlas. Bioinformatics 29(10), 1333–1340 (2013)
2. Almeida, J.S., Dress, A., Kühne, T., Parida, L.: ICT for Bridging Biology and Medicine. Dagstuhl Manifestos 3(1), 31–50 (2014)
3. Park, S., Parwani, A.V., Aller, R.D., Banach, L., Becich, M.J., Borkenfeld, S., et al.: The history of pathology informatics: A global perspective. Journal of Pathology Informatics 4 (2013)
4. Almeida, J.S., Iriabho, E.E., Gorrepati, V.L., Wilkinson, S.R., Grüneberg, A., Robbins, D.E., et al.: ImageJS: personalized, participated, pervasive, and reproducible image bioinformatics in the web browser. Journal of Pathology Informatics 3 (2012)
5. Almeida, J.S., Grüneberg, A., Maass, W., Vinga, S.: Fractal MapReduce decomposition of sequence alignment. Algorithms for Molecular Biology 7, 12 (2012)

Table of Contents

Data Integration Platforms and Applications

An Asset Management Approach to Continuous Integration of Heterogeneous Biomedical Data 1
Robert E. Schuler, Carl Kesselman, and Karl Czajkowski

Mining Linked Open Data: A Case Study with Genes Responsible for Intellectual Disability 16
Gabin Personeni, Simon Daget, Céline Bonnet, Philippe Jonveaux, Marie-Dominique Devignes, Malika Smaïl-Tabbone, and Adrien Coulet

Data Integration between Swedish National Clinical Health Registries and Biobanks Using an Availability System 32
Ola Spjuth, Jani Heikkinen, Jan-Eric Litton, Juni Palmgren, and Maria Krestyaninova

Biodiversity Data Management

Data Management Experiences and Best Practices from the Perspective of a Plant Research Institute 41
Daniel Arend, Christian Colmsee, Helmut Knüpffer, Markus Oppermann, Uwe Scholz, Danuta Schüler, Stephan Weise, and Matthias Lange

A Semantic Web Faceted Search System for Facilitating Building of Biodiversity and Ecosystems Services 50
Marie-Angélique Laporte, Isabelle Mougenot, Eric Garnier, Ulrike Stahl, Lutz Maicher, and Jens Kattge

Handling Multiple Foci in Graph Databases 58
Jaudete Daltio and Claudia Bauzer Medeiros

Ontologies and Visualization

Completing the is-a Structure of Biomedical Ontologies 66
Zlatan Dragisic, Patrick Lambrix, and Fang Wei-Kleiner

Annotation-Based Feature Extraction from Sets of SBML Models 81
Rebekka Alm, Dagmar Waltemath, Olaf Wolkenauer, and Ron Henkel

REX – A Tool for Discovering Evolution Trends in Ontology Regions . . . 96
Victor Christen, Anika Groß, and Michael Hartung

Towards Visualizing the Alignment of Large Biomedical Ontologies 104
Catia Pesquita, Daniel Faria, Emanuel Santos, Jean-Marc Neefs, and Francisco M. Couto

Linked Data and Query Processing

Discovering Relations between Indirectly Connected Biomedical Concepts 112
Dirk Weissenborn, Michael Schroeder, and George Tsatsaronis

Exploiting Semantics from Ontologies and Shared Annotations to Partition Linked Data 120
Guillermo Palma, Maria-Esther Vidal, Louiqa Raschid, and Andreas Thor

ConQuR-Bio: Consensus Ranking with Query Reformulation for Biological Data 128
Bryan Brancotte, Bastien Rance, Alain Denise, and Sarah Cohen-Boulakia

An Introduction to the Data Retrieval Facilities of the XQt Language for Scientific Data 143
Javad Chamanara and Birgitta König-Ries

Author Index 151

An Asset Management Approach to Continuous Integration of Heterogeneous Biomedical Data

Robert E. Schuler, Carl Kesselman, and Karl Czajkowski

Information Sciences Institute, University of Southern California
{schuler,carl,karlcz}@isi.edu

Abstract. Increasingly, advances in biomedical research are the result of combining and analyzing heterogeneous data types from different sources, spanning genomic, proteomic, imaging, and clinical data. Yet despite the proliferation of data-driven methods, tools to support the integration and management of large collections of data for purposes of data driven discovery are scarce, leaving scientists with ad hoc and inefficient processes. The scientific process could benefit significantly from lightweight methods for data integration that allow for exploratory, incrementally refined integration of heterogeneous data. In this paper, we address this problem by introducing a new asset management based approach designed to support continuous integration of biomedical data. We describe the system and our experiences using it in the context of several scientific applications.

1 Introduction

Biomedical advances are driven at the intersections of data: combining imaging, genetic, clinical, and other sources in cross cutting analytic methods. It is not uncommon to see a dozen different types of biomedical data, spanning genetics, multiple imaging modalities, proteomics, and clinical elements, being used in a single exploration or discovery process, each data with its own unique representation. A logical prerequisite for analysis is that the necessary data has been integrated into a formal, standard, clean, consistent, accessible, and linked representation prior to analysis. However, the vast majority of scientific data in daily use does not exist in a manner that meets even a few, if any, of the above characteristics. It is widely understood that "data wrangling" is often the most resource intensive activity in data analysis – a time-consuming process of data selection, transformation, and cleansing. All too often it is only at the very end of the scientific discovery process, while preparing data for submission into online repositories that data is integrated, organized, and annotated according to overarching standard dictionaries, ontologies, and open formats. Instead throughout most of the scientific processes, the vast majority of research data exist in semistructured, locally coded sources and formats. One consequence is that scientists often spend significant amounts of their research time managing, combining, and manipulating data, with self-reported values of 90% being common [1]. With the increasing proliferation of large biomedical data (e.g. big data), the problems will only grow worse.

H. Galhardas and E. Rahm (Eds.): DILS 2014, LNBI 8574, pp. 1–15, 2014.

In spite of this situation, it is remarkable that there is little support for scientists to integrate and organize data for purposes of exploration, analysis, and ultimate publication. Shared file systems with data organized in directory hierarchies and with metadata coded into "meaningful" file names are the common practice. Idiosyncratic methods are used to capture and unify pertinent metadata such as phenotype, experiment details, preparation methods, and quality control flags. All too often spreadsheets, which are hard to maintain and offer limited query ability, are the preferred means of describing and tracking data. Cloud based services such as Dropbox and Google Drive can provide some relief with respect to sharing, but do little to address the fundamental issues of integration and organization.

This paper presents a system that fills this gap by enabling continuous integration of heterogeneous biomedical data throughout the research and discovery lifecycle. This "pay-as-you-go" approach, influenced by the concept of Dataspaces [2], uses a process of incremental refinement to promote flexible, use-case driven data integration, which can mesh with the requirements of the data task at hand. To maximize the use of these methods by scientists, we have incorporated them into a digital asset management system for biomedical data (BDAM) that resembles cloud based tools and services with which investigators are already familiar. The structure of this paper is as follows. In Section 2, we introduce the concept of digital asset management and continuous integration to build a unified view over heterogeneous life science data collections. In Section 3, we discuss related work. Section 4, presents design of a biomedical digital asset management system whose application in a range of use cases is discussed in Section 5. Finally we describe future plans and conclusions .

2 Asset Management Method for Continuous Integration

Digital asset management (DAM) "consists of management tasks and decisions surrounding the ingestion, annotation, cataloguing, storage, retrieval and distribution of digital assets" [3]. DAM systems are designed to streamline free-form "creative" processes rather than enforce predefined business processes. DAM is nearly ubiquitous for many varieties and applications of data, from text and document management to multimedia to specialized systems for marketing and the web. For example, DAM systems for photo management like iPhoto or Picasa will discover and catalog digital images on one's hard disk drive, extract metadata from the imported media, cleanse (fix or add missing) metadata, allow user annotations (typically in the form of tags), organize pictures into virtual collections (i.e. photo albums), support browsing and search, support data export for data manipulation by external photo editing tools, and support publication for cloud based sharing or printing by online services.

Surprisingly, in spite of the fact that there would seem to be a good alignment between the data management requirements for biomedical discovery and the functions provided by DAM systems, DAM approaches have not been generally applied to biomedical data management. Building on the success of DAM in other creative fields, we claim that an approach to data integration that assists scientists throughout the research lifecycle based on a biomedical digital asset management system (BDAM) would significantly streamline the process of data driven scientific discovery in the life sciences. However, it is also the case that simply applying an existing

DAM technology will be insufficient to meet the needs of life sciences where the data are both large and significantly more diverse.

2.1 Continuous Integration of Biomedical Data

As discussed above, a core requirement of BDAM is the ability to manage a heterogeneous collection of asset types, each with their own characteristics, descriptive metadata, and storage representation (i.e. file format). Within the overall function of asset management, one can take a broad perspective of what it means to provide integration based on the management operations under consideration: i.e. search, organization, or export for analysis. We may limit integration to the descriptive metadata associated with each asset (or collection of assets) or we may want to provide a uniform rendering of the structure of the asset itself. For example, assets of diverse types can be integrated simply by being grouped into logical collections, assets may be collected based on shared common attributes and criteria (i.e. faceted search [4]), or the underlying structure of the assets themselves may be transformed, transcoded, or reformatted into a uniform representation.

Common approaches to data integration, which depend on tight semantic integration of traditional database Extraction-Transformation-Loading (ETL), upfront semantic alignment and schema mapping (e.g., query mediation [5]) are problematic when the descriptive data is not known beforehand, or may change during the discovery process, which is often the case in life sciences application. Consequently, an incremental model which assumes that metadata is incomplete [6] or evolving [7], that assumes loose semantic integration, no upfront semantic alignment, loose administrative proximity, and loose consistency with sources will have broader applicability than a non-incremental approach. Building on the axiom of “integrate early and often” approaches such as Dataspaces [2] or MAD [8] seek to accelerate the use of data by deferring integration until required. We embrace this model as a core aspect of BDAM by providing functions for editing, augmenting, and refining metadata descriptions incrementally over the lifetime of the discovery process. This is not to say that established models cannot be used, even in early phases of data use. With BDAM we take a hybrid approach where structured metadata is ingested into the system and augmented with incrementally defined descriptions.

3 Related Work

Digital repository systems (e.g. DSpace [9]) provide capabilities aimed at long-term preservation and archiving of scholarly works. They are primarily concerned with document management (Word, PDF, JPEG, etc.), whereas a DAM system for life sciences must support diverse biomedical file formats and very large file sizes and overall file volumes. Digital repositories support publication and archiving, thus they should be viewed as an endpoint for the scientific data assets produced by researchers. Plale et al [7] have proposed the SEAD Virtual Archive for federating institutional repositories along with automated workflows to assist researchers in the data publication process. The asset management approach proposed here takes this a significant step further by pushing deeper and earlier into the scientific discovery process so that data curation is not an overhead but an integral part of the discovery processes.

SQLShare [1] is a system that has many elements in common with the BDAM catalog including the concepts of schema evolution and incremental refinement. However, SQLShare differs from our work in several significant ways. It focuses on SQL as the primary interface by which users interact and assumes that the data of interest is primarily stored in the SQLShare database. Metadata catalogs, such as Globus Metadata Catalog Service were proposed [10], with an extensible schema as a general purpose tool to support data management in e-sciences. The asset management approach argues that metadata catalogs must be coupled with semi-automated methods for metadata ingest and complementary asset management services.

Picture Archiving and Communications Systems (PACS) based on the DICOM standard for medical imaging interoperability offer clinical image management services with interfaces to store, query, and retrieve radiology images. Related are research systems such as XNAT [11]. These systems, however, are focused almost exclusively on radiology imaging rather than other imaging modalities or data types, and they do not offer schema evolution, as described later.

Finally, storage management systems, including SRM [12], and iRODS [13], provide facilities for lower-level data storage operations and storage resource management. They generally operate on data at a semantically lower level than digital asset management and offer limited facilities for metadata management.

4 BDAM Design and Implementation

The core elements of any DAM system include: 1) a catalog for tracking, managing and organizing assets, 2) ingest methods for incorporating data assets into the system

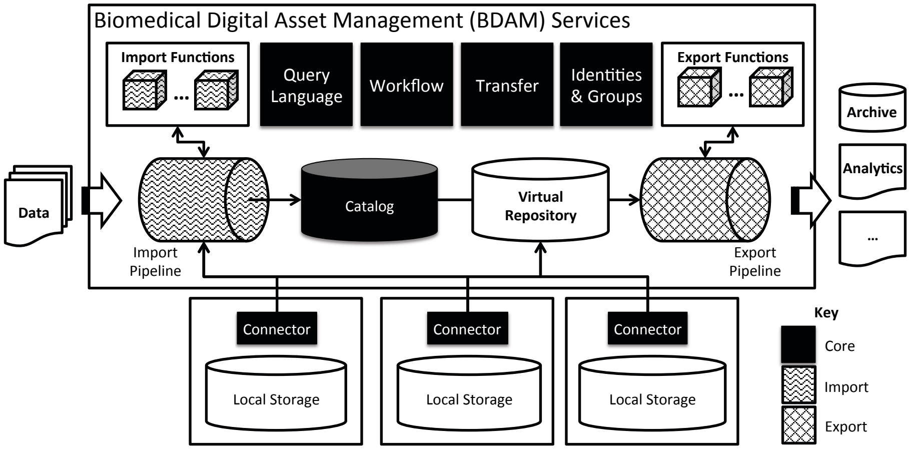

Fig. 1. The architecture of a biomedical digital asset management service. BDAM services are loosely coupled via connectors to local storage and they facilitate import and export pipelines with extensible functions.

and extracting basic descriptive metadata, 3) storage services for storing and moving assets, and 4) methods for extracting assets from the system for analysis and publication. Pervasive across all these functions must be methods for specifying and enforcing policy for access and use. The relationship between these functions is shown in Fig 1.

4.1 Design Requirements

The heterogeneity of data assets and scientific processes means that the semantic and syntactic models for metadata and data are often not known a priori. Furthermore, the metadata characterizing a particular asset and its relationship to the research or discovery task at hand is not a simple functional product of the data content but may vary depending on the research questions being posed and the kind of data discovery that will be performed. We adopt a *relaxed consistency* model in which a data asset continually evolves, rather than entering the system fully formed and with all metadata predetermined. We allow *incremental refinement* of content and schema, throughout scientific discovery.

The disparate sources of data assets are not always under one administrative domain. Scientific discovery may involve assets located in a combination of local, enterprise and cloud based storage. In many cases restricted access data covered by Institutional Review Boards, government regulations such as Health Insurance Portability and Accountability Act, and other Data Use Agreements may not permit the use of clouds for storage of sensitive data. We adopt a hybrid design [14] with *loose coupling*, in which core components are operated in a software-as-a-service (SaaS) platform while user data may reside in local storage services (see Fig. 1). The complexities of managing the core services are reduced by operating in a hosted environment, while institutional data access controls are preserved and storage costs are lowered.

In addition to schema evolution, the BDAM must support *schema introspection*. Given the dynamic nature of schema evolution, the applications and user interfaces must be able to inspect the catalog's schema and present interfaces for the user to query and manipulate content. Often, useful metadata that characterizes data assets may contain private information, and it is not enough to assume access control for data assets while having unrestricted access to metadata. A BDAM must support *fine-grain access control* to restrict access to metadata about specific assets (e.g. rows or resources), to attributes (e.g. columns or property types) of the any asset, or to whole collection of metadata (e.g. tables or graphs). Data storage services should support complementary access control to the data.

Finally, if a BDAM is offered as a shared service it must support *multi-tenancy* to allow each scientific application to operate at its own pace and with its own content and access policies. All these design characteristics (loose coupling, relaxed consistency, incremental refinement, fine-grain access control, and multi-tenancy) complement one another. The BDAM is able to capture the evolving state of the scientific discovery process as data assets are acquired, summarized, queried, processed, and analyzed by researchers.

4.2 Data Catalog

The BDAM data catalog allows individual data assets or other relevant resources to be recorded along with meaningful metadata descriptions. As one of the loosely coupled components of the BDAM, the catalog may receive input from multiple sources including direct, user-authored metadata and machine-driven metadata extraction tools. These catalog contents can be browsed or searched to find assets matching certain criteria. The catalog schema can be queried and amended as per our schema evolution and introspection design requirements. Metadata concepts must be defined before first use, but these definitions can be incrementally added to the running catalog at any point during its operation.

In keeping with the SaaS model, interactions with the catalog are via a RESTful web services protocol. The catalog contains metadata records as resources which can be manipulated by the client. The defined interface includes functions for retrieving and amending the metadata schema; creating, destroying, updating, and retrieving whole metadata records; updating or retrieving individual metadata properties for specific records; or performing queries of the records by metadata criteria and associations to other contextual records. The metadata update and retrieval interfaces also allow bulk operations to efficiently manipulate many records in a single request.

We have developed and evaluated two distinct catalog implementations, both presenting a web service access protocol on top of a relational database management system (RDBMS). Based on the widespread appeal of graph-base query in data integration, we initially explored sparse data storage models with a graph-based query interface, which we called Tagfiler (described later). However, we found that in practice, many investigations use a handful of dominant resource models where many assets were annotated with the same subset of metadata concepts. In such an environment, it is desirable to use a more compact representation of metadata. Consequently, we developed an alternative catalog interface which supports more structured modeling of data. *ERMrest*, a portmanteau of ERM (Entity Relationship Model) and REST (REpresentational State Transfer) exposes a table-like concept of typed entities with type-specific properties and is tuned for dense metadata by storing entities and their properties as rows in conventional tables.

In both ERMrest and Tagfiler, the catalog model exposes not only individually named metadata records, but also complex record sets. Both also support complex query patterns where assets can be found not only by their direct metadata but also by their relationships to other matching assets. In each, the web access model defines a structured naming scheme (i.e. URI) to denote computed record sets based on attribute matching patterns. However, the two catalogs do not implement the same naming scheme. In Tagfiler, the naming scheme was tuned for encoding arbitrary graph-query patterns, where one computed set of assets could be derived from another by traversing an arbitrarily chosen linking property.

In ERMrest the relationships between entity types are also captured in the typed schema, corresponding to the underlying RDBMS concept of foreign-key references between tables. ERMrest defines a compact URI naming scheme to traverse such linked entities as a psuedo-hierarchy of related entity sets. A URI denoting one set of