Multiscale Dataspace for Organism-centric Analysis

Matheus Silva Mota[1], **Júlio Cesar dos Reis**[1], **Sandra Goutte**[2], **André Santanchè**[1]

[1]Institute of Computing [2]Institute of Biology – UNICAMP
Campinas – SP – Brazil

mota@ic.unicamp.br, julio.dosreis@ic.unicamp.br

froggologist@gmail.com, santanche@ic.unicamp.br

Abstract. *Biologists increasingly need a unified view to understand and discover relationships among data elements scattered along data sources with different levels of heterogeneity. Existing approaches usually adopt ad-hoc* heavyweight *integration strategies, requiring a costly upfront effort involving a monolithic chain of steps to handle specific formats/schemas, with low or no reuse. This article proposes an original framework based on scales aligned with the dataspaces on demand integration principle. Scales systematize and encapsulate integration in discrete steps, fulfilling the dynamicity of the process through reuse of previous scales and localized customization. Although the proposed framework can be extended to several scenarios, this work focuses on the biology domain addressing the organism-centric analysis scenario.*

1. Introduction

Data-centric domains as biology are increasingly adopting different systems to produce, store and analyze datasets regarding specific processes and aspects of biological organisms – e.g., experiments, descriptions, collections, simulations, etc. However, heterogeneity hampers the exploration of knowledge across systems and research groups in an integrated way [Hey et al. 2009]. Therefore, integration is a key challenge since providing specialized and big picture-like views of data may offer new perspectives and insights to researchers [Elsayed and Brezany 2010].

This work focus on a specific integration approach known as Dataspaces. It advocates the advantages of an on demand lightweight integration, to comply with the dynamicity of modern environments, against the classic heavyweight upfront strategies. One of the advantages of on demand integration is the ability of readily shaping the final product according to present needs. A problem with on demand integration addressed in this investigation refers to the long chain of steps from source to target. In one extreme, biologists want to treat knowledge in a conceptual level, handling data in an integrated fashion. In the other extreme, there are several problem-relevant heterogeneous data sources, comprising files, DBs, ontologies, etc. Between both extremes, there is a spectrum of intermediary integration steps.

In this article, we propose an approach named *LinkedScales*, which aims at splitting such integration steps as discrete scales. Each scale encompasses common aspects and routines related to a specific integration step. *LinkedScales* objective is going from a source-related lower scale, to a user-focused higher scale. Inspired by the layered software architecture, each scale offers to the immediate upper scale a pre-agreed model (interface), encapsulating heterogeneities of lower scales handled until its stage.

We demonstrate the applicability of our proposal in the biological domain. In such dynamic context, reuse becomes a challenge, since on demand solutions usually rely on ad-hoc solutions, implementing the entire integration chain. The encapsulation of scales in *LinkedScales* enables to customize only algorithms of a specific scale, reusing the remaining of the chain. In lower scales, we depart from myriad available heterogeneous sources. The upper scale enables to tailor the model according to specific needs, i.e., the integration model fits to the user needs, instead of the opposite.

This paper is organized as follows: Section 2 describes and exemplifies the addressed problem. Section 3 presents our multiscale integration proposal and its utility for biological data. In Section 4 we wrap up the article with some conclusions.

2. Challenges on organism-centric analysis

Organism-centric analysis refers to an usual approach conducted by biologists in which organisms – i.e., species or taxonomic groups – are the central focus of the analysis and data are integrated around them. A common task faced by biologists conducting an organism-centric research is the construction of "views" of data, we call here *profiles* [Washington et al. 2009]. A profile varies according to the focus of interest, but they can be seen as a subset of descriptive data of organisms selected for a research [Hedges 2002]. The construction of such profiles involves combining data usually fragmented in heterogeneous sources, requiring further efforts from biologists to collect and combine pieces coming from multiple repositories and several files with different formats.

Consider the example of profile illustrated in Figure 1, defined by biologists interested in validating hypotheses regarding the evolution of "deafness" in frogs. Aiming at understand why distant phylogenetic groups of frogs lack middle ear structures, biologists want to gather together as profiles data regarding morphological traits, habitat, reproduction mode, acoustics and phylogenetic trees of several species. Morpho-anatomical data would be required to examine whether miniaturisation in frogs lead to the loss of ear structures, while acoustic data would allow testing the co-evolution of mutism and deafness, etc. Based on such profiles, biologists might compare organisms in a systematic way and investigate conditions and associations related with the hypotheses.

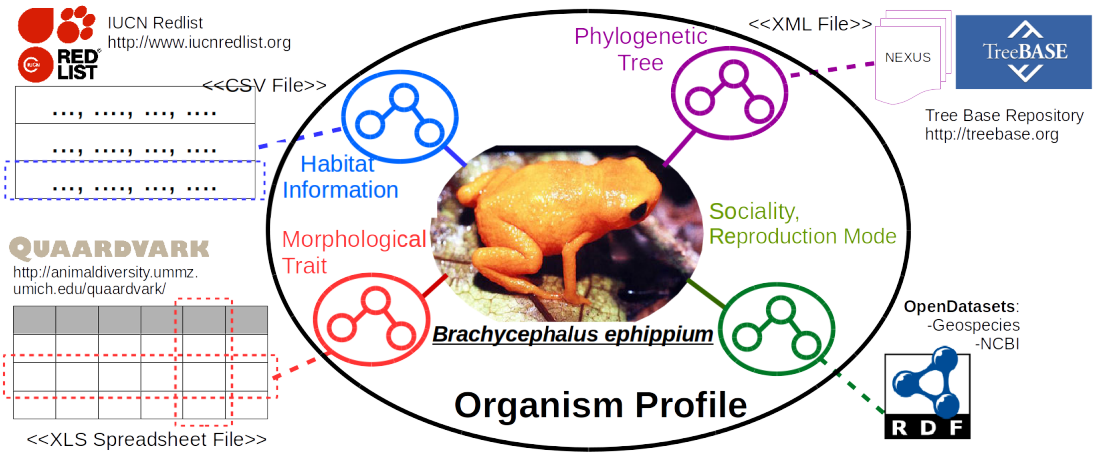

Figure 1. Profile integrating characteristics scattered across several sources

Phylogenetic data for the target species of the genus *Brachycephalus* (shown in Figure 1) can be found within the TreeBASE[1] repository – where scientists share their experimental data files – as a XML/Nexus file. It contains the phylogenetic tree reconstructed from DNA sequences from a study. Moreover, conservation data from IUCN Red List[2] in CSV format contains data regarding the species habitat and several phenotipic data can be found from Quaardvark System[3] in Excel format.

In this scenario (Figure 1), biologists spend a lot of time "cutting and pasting" data from each of the sources and organizing them in spreadsheets before any analysis. On the other hand, a systematic integration approach requires several steps of integration, due to the different types of heterogeneity, i.e., different formats (CSV, Excel, Nexus), different structures (tables, trees), different schemas, etc. Therefore, the combination of different types of datasets may prove challenging, and the integration of missing data often result in a drastic data trimming and the partial use of the data available. Furthermore, such research has an intrinsic dynamism. For instance, biologists may discover during the research that other characteristics must be taken into account, which might require further efforts to reflect the new requirements on the profiles.

3. Multiscale dataspace as a basis for organism-centric analysis

There is a modern tendency towards progressive integration [Halevy et al. 2006] – known as *dataspace* – that goes in the opposite direction of the classical ad-hoc heavy weight integration [Singh and Jain 2011]. Possible changes on the profiles during the research and the heterogeneity of formats, models and schemas on the sources make the dataspace approach a powerful alternative for integration of data during an organism-centric research. A dataspace may offer progressive snapshots to the biologists. – i.e., profiles are "filled" on demand, but faster than manually, not hampering initial analysis of data as it is being integrated, also better accommodating changes on the variables and sources. Literature proposes approaches for achieving and maintain a dataspace [Singh and Jain 2011]. Nevertheless, most of them treat integration as a monolithic task, and fail considering that there is a common underlying pipeline in the integration chain addressing different aspects of heterogeneity, which produces intermediary reusable assets.

LinkedScales refers to a framework aiming at bringing the proposal of multiscale to the data integration chain, systematizing and encapsulating the data regarding integration steps as scales. It starts by transforming all data sources into graphs and takes advantage of the flexibility of graph structures to logicaly represent multiple scales. This allows formal operations within and across the scales as graphs transformation procedures. *LinkedScales* systematically defines – based on previous experiences on data integration [Mota and Medeiros 2013, Bernardo et al. 2013] – a fix initial set of scales, where each scale focuses in a different level of integration and its respective abstraction. Steps are interconnected in the graph, supporting traceability among the scales – i.e., it is possible to "track" the transformations and the source that produced the data within scales.

Figure 2 depicts an overview of the *LinkedScales* framework applied to an organism-centric analysis. It presents four different scales of abstraction aiming at going from the sources (lower scales, containing more details about format and structure) to a conceptual scale (less details of format and structure, and focus on organism-centric

[1]http://treebase.org, [2]http://www.iucnredlist.org, [3]http://animaldiversity.ummz.umich.edu/quaardvark

concepts). From bottom to top, the scales are: (i) *Physical Scale*, (ii) *Logical Scale*; (iii) *Description Scale*; and (iv) *Conceptual Scale*. Further scales can appear on top of the conceptual scale to define additional domain-related views.

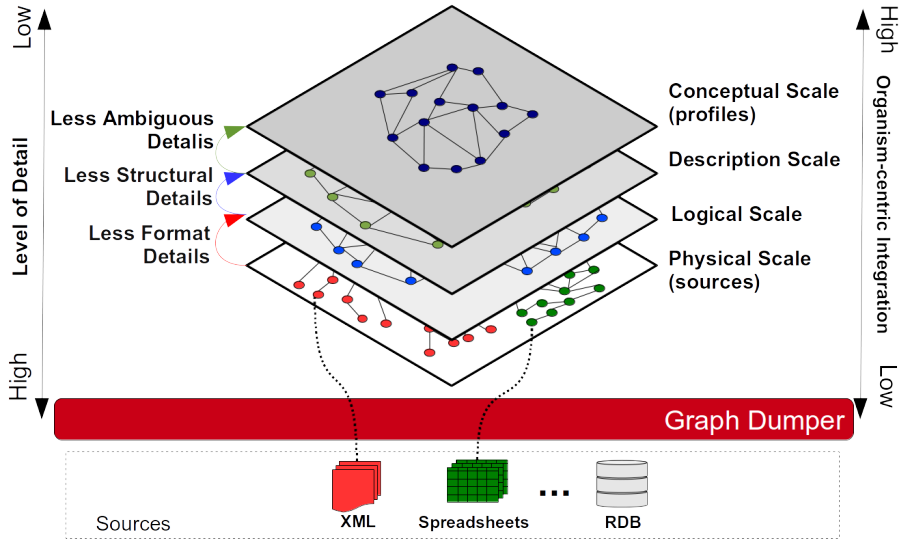

Figure 2. *LinkedScales* framework applied to an organism-centric analysis

The lowest scale of Figure 2 – the ***Physical Scale*** – aims at representing the different data sources in their original physical format. It is the lowest-level raw content containing a format representation of the data sources with addressable/linkable component items. The original structure and content of the underlying data sources is reflected in a graph, as far as possible. The role of this scale – in an incremental integration process – concerns making explicit and linkable original data within sources.

Even though our proposal can be extended to other file formats, we are currently focusing only on a set of formats defined by biologists as the most relevant for their work (discussed in Section 2), i.e., spreadsheets (XLS, XLSX, ODS), HTML tables, CSV files, XML files and textual documents. We have developed a graph API for ETL named *2graph*[2] that might support biologists building profiles. The API is represented as the "*Graph Dumper*" element in Figure 2 and stores data within the Neo4j graph database.

Based on experiences of a previous work that explores a homogeneous representation model for textual documents independently of formats [Mota and Medeiros 2013], the next scale proposed is the ***Logical Scale***. It offers a common view to data inside similar or equivalent structural models represented in the previous scale. Tables and hierarchical documents are examples of structural models present in the sources containing data regarding organisms. In the previous scale, differences might exist in the representation of a table within a PDF, a table from a spreadsheet and a table within a HTML file, since they preserve specificities of their formats. Within the *Logical Scale*, format specificities disappear and the three tables are represented alike since they refer to the same structural model. This leads to a homogeneous approach to process tables, independently of the way that tables are represented in their original specialized formats.

[2]Available at http://www.lis.ic.unicamp.br/ matheus/projects/2graph

The ***Description Scale*** emphasizes the content (e.g., labels of tags within a XML or values in spreadsheet cells) and their relationships. Since models represent relation among data elements in different ways – e.g., a row in a table can represent data concerning the same entity, while hierarchy relies on aggregations – the *Description Scale* reduces all models to a single unified one, to shift the focus towards the descriptive content, avoiding heterogeneous models concerns. The unified model selected for this scale is based in the simple triple <resource, property, value>, which is usual in several metadata standards as RDF. A triple-extraction algorithm can be applied and it is currently under investigation.

Although biologists still cannot handle data from previous scales in a conceptual and more integrated fashion, the *Description Scale* can be helpful to them, as it already allows some preliminary and meaningful analysis. For instance, spreadsheets regarding morphological traits usually adopts a cross-sheet way of organization. Such organization hampers an unified view of the traits of an organism, requiring more efforts from the biologists when conducting any initial analysis.

The highest scale of Figure 2 is the ***Conceptual Scale***. This scale accounts for integrating data of the lower scale in a semantic level, exploiting relationships between nodes to discover and to make explicit as ontologies the latent semantics in the content. At this level, it is possible to deduplicate common entities and to infer their semantics – e.g., instances of classes in ontologies – and their properties. We also consider that predefined ontologies can be straightly mapped to this scale, to be related to the inferred entities.

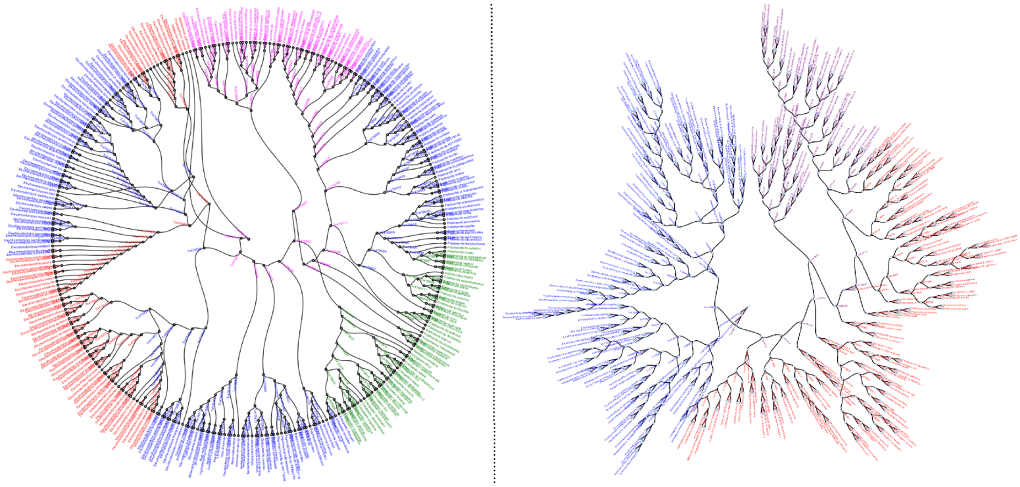

Figure 3. Example of visualization of the Description Scale

At this stage of the investigation, *LinkedScales* enables to integrate XML files containing phylogenetic trees (from the TreeBase repository) with spreadsheets and CSV files regarding morphological traits (maintained by biologists). Based on the homogeneous models produced for the files in the *Logical Scales* (after being represented as a raw-format in the *Physical Scale*), species names mentioned on the tree and species names mentioned on the tables are linked using a simple string match. Figure 3 illustrates a visualization of output results corresponding to the initial outcome from the *Description Scale*. It shows the species following the phylogenetic tree provided by the XML file aggregated (colors) according to the tables in which the species are mentioned.

4. Conclusion

A significant part of the biological research works in an organism-centric perspective, which usually requires combining data regarding distinct aspects of organisms. However, relevant data is typically scattered among heterogeneous sources with different formats, structures and schemas, hampering the combination of data across sources to perceive information meaningfully and to systematically compare organisms. In this paper, we propose an original framework, named *LinkedScales*, based on the multiscale integration approach. Our proposal allows an homogeneous perspective of data in each scale, encapsulating details about heterogeneities. We showed the potential benefits of *LinkedScales* to reach organism profiles. Future work involves formalizing the elements of the framework and the transformation operations as well as conducting an experimental evaluation.

Acknowledgments

Work partially financed[3] by CNPq (#141353/2015-5), FAPESP (#2014/14890-0), Microsoft Research FAPESP Virtual Institute (NavScales project), FAPESP/Cepid in Computational Engineering and Sciences (#2013/08293-7), CNPq (MuZOO Project), INCT in Web Science, FAPESP-PRONEX (eScience project), and individual grants from CAPES.

References

Bernardo, I. R., Mota, M. S., and Santanchè, A. (2013). Extracting and semantically integrating implicit schemas from multiple spreadsheets of biology based on the recognition of their nature. *Journal of Info. and Data Manag.*, 4(2):104.

Elsayed, I. and Brezany, P. (2010). Towards large-scale scientific dataspaces for e-science applications. In *Database Systems for Advanced Applications*, pages 69–80. Springer.

Halevy, A., Franklin, M., and Maier, D. (2006). Principles of dataspace systems. In *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART*, PODS '06, pages 1–9, New York, NY, USA. ACM.

Hedges, S. B. (2002). The origin and evolution of model organisms. *Nature Reviews Genetics*, 3(11):838849.

Hey, T., Tansley, S., and Tolle, K., editors (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, Washington.

Mota, M. and Medeiros, C. (2013). Introducing shadows: Flexible document representation and annotation on the web. In *Proc. of Data Engineering Workshops (ICDEW), IEEE 29th ICDE*, pages 13–18.

Singh, M. and Jain, S. (2011). A survey on dataspace. In Wyld, D., Wozniak, M., Chaki, N., Meghanathan, N., and Nagamalai, D., editors, *Advances in Network Security and Applications*, volume 196 of *Communications in Computer and Information Science*, pages 608–621. Springer Berlin Heidelberg.

Washington, N. L., Haendel, M. A., Mungall, C. J., Ashburner, M., Westerfield, M., and Lewis, S. E. (2009). Linking human diseases to animal models using ontologybased phenotype annotation. *PLoS biology*, 7(11):e1000247.

[3]The opinions expressed in this work do not necessarily reflect those of the funding agencies.