Conceiving a Multiscale Dataspace for Data Analysis

Matheus Silva Mota[1], **André Santanchè**[1]

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

{mota,santanche}@ic.unicamp.br

Abstract. *A consequence of the intensive growth of information shared online is the increase of opportunities to link and integrate distinct sources of knowledge. This linking and integration can be hampered by different levels of heterogeneity in the available sources. Existing approaches focusing on* heavyweight *integration – e.g., schema mapping or ontology alignment – require costly upfront efforts to handle specific formats/schemas. In this scenario, dataspaces emerge as a modern alternative approach to address the integration of heterogeneous sources. The classic heavyweight upfront one-step integration is replaced by an incremental integration, starting from* lightweight *connections, tightening and improving them when benefits worth such effort. Based on several previous work on data integration for data analysis, this work discusses the conception of a multiscale-based dataspace architecture, called LinkedScales. It departs from the notion of integration-scales within a dataspace, and defines a systematic and progressive integration process via graph-based transformations over a graph database. LinkedScales aims to provide a homogeneous view of heterogeneous sources, allowing systems to reach and produce different integration levels on demand, going from raw representations (lower scales) towards ontology-like structures (higher scales).*

1. Introduction and Motivation

From science to business, several domains are facing a huge increase in the amount of available data and the growth of the data heterogeneity (in various levels). In parallel, opportunities may emerge from the exploitation of the increasing volume of connections among multidisciplinary data [Hey et al. 2009].

Domains like biology are increasingly becoming data-driven. Although they adopt different systems to produce, store and search their data, biologists increasingly need a unified view of these data to understand and discover relationships between low-level (e.g., cellular, genomic or molecular level) and high-level (e.g., species characterization, macro-biomas etc.) biological information among several heterogeneous and distributed sources. Therefore, integration becomes a key factor in such data-intensive and in multidisciplinary domains; the production and exploitation of connections among independent data-sources become essential [Elsayed and Brezany 2010]. Besides integration, challenges like provenance, visualization and versioning are experienced by domains that handle large, heterogeneous and cross-connected datasets [Heath and Bizer 2011].

In order to integrate available sources, classical data integration approaches, found in the literature, usually require an up-front effort related to schema recognition/mapping in an all-or-nothing fashion [Halevy et al. 2006a]. On demand integration of distinct and heterogeneous sources requires ad hoc solutions and repeated effort from specialists [Franklin et al. 2005].

Franklin et. al propose the notion of *dataspaces* to address the problems mentioned above [Franklin et al. 2005]. The dataspace vision aims to provide the benefits of the classical data integration approach, but via a progressive "pay-as-you-go" integration [Halevy et al. 2006a]. They argue that linking lots of "fine-grained" information particles, bearing "little semantics", already bring benefits to applications, and more links can be produced on demand, as *lightweight* steps of integration.

Related work proposals address distinct aspects of dataspaces. Regarding the architectural aspect, each work explores a different issue of a dataspace system. Among all efforts, no dominant proposal of a complete architecture has emerged until now. We observed that, in a progressive integration process, steps are not all alike. They can be distinguished by interdependent roles, which we organize here as abstraction layers. They are materialized in our LinkedScales, a graph-based dataspace architecture. Inspired by a common backbone found in related work, LinkedScales aims to provide an architecture for dataspace systems that supports progressive integration and the management of heterogeneous sources.

LinkedScales takes advantage of the flexibility of graph structures and proposes the notion of scales of integration. Scales are represented as graphs, managed in graph databases. Operations become transformations of such graphs. LinkedScales also systematically defines a set of scales as layers, where each scale focuses in a different level of integration and its respective abstraction. In a progressive integration, each scale congregates homologous lightweight steps. They are interconnected, supporting provenance traceability. Furthermore, LinkedScales supports a complete dataspace lifecycle, including automatic initialization, maintenance and refinement of the links.

This paper discusses the conceiving of the LinkedScales architecture and is organized as follows. Section 2 discusses some concepts and related work. Section 3 introduces the LinkedScales proposal, also discussing previous work and how such experiences led to the proposed architecture. Section 4 presents previous work on data integration and discusses how such experiences are reflected in current proposal. Finally, Section 5 presents some conclusions and future steps.

2. Related Work

2.1. The Classical Data Integration

Motivated by such increasingly need of treating multiple and heterogeneous data sources, data integration has been the focus of attention in the database community in the past two decades [Hedeler et al. 2013]. One predominant strategy is based on providing a virtual unified view under a global schema (GS) [Kolaitis 2005]. Within GS systems, the data stay in their original data sources – maintaining their original schemas – and are dynamically fetched and mapped to a global schema under clients' request [Lenzerini 2002, Hedeler et al. 2013]. In a nutshell, applications send queries to a mediator, which maps them into several sub-queries dispatched to wrappers, according to metadata regarding capabilities of the participating DBMSs. Wrappers map queries to the underlying DBMSs and the results back to the mediator, guided by the global schema. Queries are optimized and evaluated according to each DBMS within the set, providing the illusion of a single database to applications [Lenzerini 2002].

A main problem found in this “classical” data integration strategy regards the big upfront effort required to produce a global schema definition [Halevy et al. 2006b]. Since in some domains different DBMSs may emerge and schemas are constantly changing, such costly initial step can become impracticable [Hedeler et al. 2013]. Moreover, several approaches focus on a particular data model (e.g., relational), while new models also become popular [Elsayed et al. 2006]. As we will present in next section, an alternative to this classical all-or-nothing costly upfront data integration strategy is a strategy based on progressive small integration steps.

2.2. The “Pay-as-you-go” Dataspace Vision

Since upfront mapping between schemas are labor intensive and scheme-static domains are rare, pay-as-you-go integration strategies have gained momentum. Classical data integration (presented in Section 2.1) approaches work successfully when integrating modest numbers of stable databases in controlled environments, but lack an efficient solution for scenarios in which schemas often change and new data models must be considered [Hedeler et al. 2013]. In a data integration spectrum, the classical data integration is at the high-cost/high-quality end, while an incremental integration based on progressive small steps starts in the opposite side. However, this incremental integration can be continuously refined in order to improve the connections among sources.

In 2005, *Franklin et. al* published a paper proposing the notion of *dataspaces*. The dataspace vision aims at providing the benefits of the classical data integration approach, but in a progressive fashion [Halevy et al. 2006a, Singh and Jain 2011, Hedeler et al. 2010]. The main argument behind the dataspace proposal is that, in the current scenario, instead of a long wait for a global integration schema to have access to the data, users would rather to have early access to the data, among small cycles of integration – i.e., if the user needs the data now, some integration is better than nothing. This second generation approach of data integration can be divided in a bootstrapping stage and subsequent improvements. Progressive integration refinements can be based, for instance, on structural analysis [Dong and Halevy 2007], on user feedback [Belhajjame et al. 2013] or on manual / automatic mappings among sources – if benefits worth such effort.

Dataspaces comprise several challenges related to the design of Dataspace Support Platforms (DSSPs). The main goal of a DSSP is to provide basic support for operations among all data sources within a dataspace, allowing developers to focus on specific challenges of their applications, rather than handling low-level tasks related to data integration [Singh and Jain 2011]. Many DSSPs have been proposed recently addressing a variety of scenarios, e.g., SEMEX [Cai et al. 2005] and iMeMex [Dittrich et al. 2009] on the PIM context; PayGo [Madhavan et al. 2007] focusing on Web-related sources; and a justice-related DSSP[Dijk et al. 2013]. As far as we know, up to date, the proposed DSSPs provide specialized solutions, targeting only specific scenarios [Singh and Jain 2011, Hedeler et al. 2009].

3. LinkedScales: A Multiscale Dataspace Architecture

The goal of LinkedScales is to systematize the dataspace-based integration process in an architecture. It slices integration levels in progressive layers, whose abstraction is inspired by the notion of scales. As an initial effort, LinkedScales strategy focuses on a specific goal on the dataspace scope: to provide a homogeneous view of data, hiding details about heterogeneous and specific formats and schemas. To achieve this goal, the current proposal does not address issues related to access policies, broadcast updates or distributed access management.

LinkedScales is an architecture for systematic and incremental data integration, based on graph transformations, materialized in different scales of abstraction. It aims to support algorithms and common tools for integrating data within the dataspaces. Integration-scales are linked, and data in lower scales are connected to their corresponding representations in higher scales. As discussed in next section, each integration-scale is based on experiences acquired in three previous experiences related to data integration.

Figure 1 shows an overview of the LinkedScales DSSP architecture, presenting, from bottom to top the following scales of abstraction. (i) *Physical Scale*, (ii) *Logical Scale*; (iii) *Description Scale*; and (iv) *Conceptual Scale*.

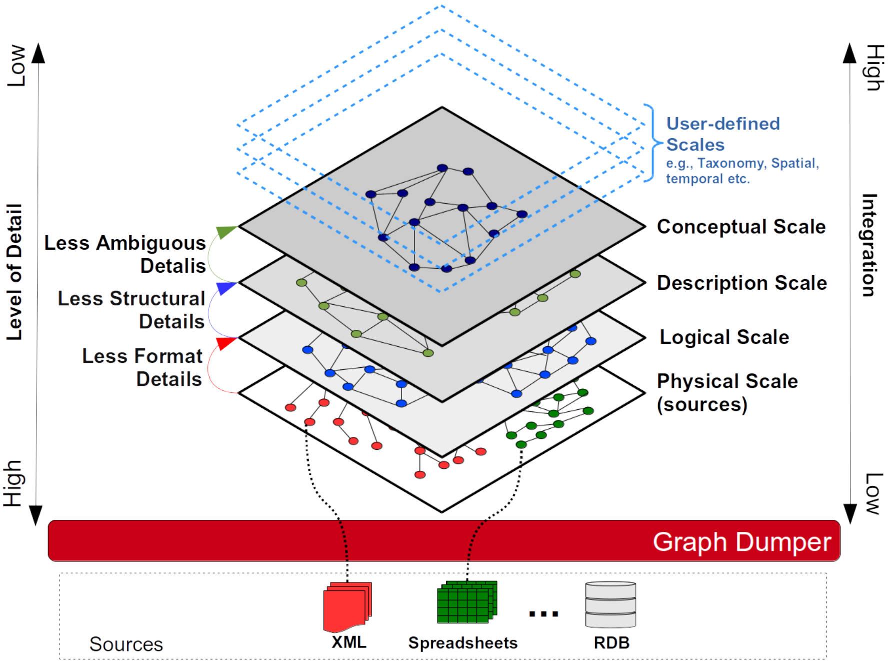


Figure 1. Overview of the LinkedScales architecture.

The lowest part of Figure 1 – the *Graph Dumper* and the *Sources* – represents the different data sources handled by our DSSP in their original format. Even though we are conceiving an architecture that can be extended to any desired format, we are currently focusing on spreadsheets, XML files and textual documents as underlying sources. Data at this level are treated as black-boxes. Therefore, data items inside the sources are still not addressable by links.

The lower scale – the *Physical Scale* – aims at mapping the sources available in the dataspace to a graph inside a graph database. This type of database stores graphs in their native model and they are optimized to store and handle them. The operations and query languages are tailored for graphs. There are several competing approaches to represent graphs inside the database [Angles 2012, Angles and Gutierrez 2008].

The *Physical Scale* is the lowest-level raw content+format representation of data sources with addressable/linkable component items. It will reflect in a graph, as far as possible, the original structure and content of the original underlying data sources. The role of this scale – in an incremental integration process – concerns making explicit and linkable data within sources. In a dataspace fashion, such effort to make raw content explicit can be improved on demand.

The *Logical Scale* aims at offering a common view to data inside similar or equivalent structural models. Examples of structural models are: table and hierarchical document. In the previous scale, there will be differences 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. In this (Logical) scale, on the other hand, the three tables should be represented in the same fashion, since they refer to the same structural model. This will lead to a homogeneous approach to process tables, independently of how tables were represented in their original specialized formats. To design the structural models of the Logical Scale we will investigate initiatives such as the OMG's[1] Information Management Metamodel[2] (IMM). IMM addresses the heterogeneity among the models behind Information Management systems, proposing a general interconnected metamodel, aligning several existing metamodels. Figure 2 presents an overview of the current state of the IMM and supported metamodels. For instance, it shows that XML and Relational metamodels can be aligned into a common metamodel.

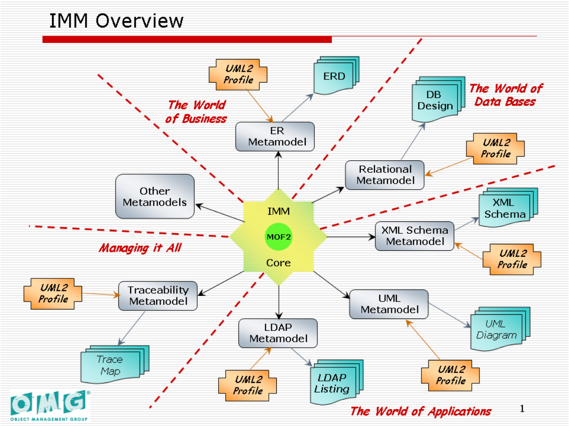

IMM Overview

Figure 2. Overview of the current state of the IMM proposal. Source: http://www.omgwiki.org/imm

In the *Description Scale*, the focus is in the content (e.g., labels of tags within a XML or values in spreadsheet cells) and their relationships. Structural information pertaining to specific models – e.g., aggregation nodes of XML – are discarded if they do not affect the semantic interpretation of the data, otherwise, they will be transformed in a relation between nodes following common patterns – for example, cells in the same

[1] Object Management Group – http://www.omg.org

[2] http://www.omgwiki.org/imm

row of a table are usually values for attributes of a given entity. Here, the structures from previous scales will be reflected as RDF triples.

The highest scale of Figure 1 is the *Conceptual Scale*. It unifies in a common semantic framework the data of the lower scale. Algorithms to map content to this scale exploit relationships between nodes of the Description Scale to discover and to make explicit as ontologies the latent semantics in the existing content. As we discuss in next section, it is possible in several scenarios to infer semantic entities – e.g., instances of classes in ontologies – and their properties from the content. We are also considering the existence of predefined ontologies, mapped straight to this scale, which will support the mapping process and will be connected to the inferred entities. Here, algorithms concerning entity linking should be investigated.

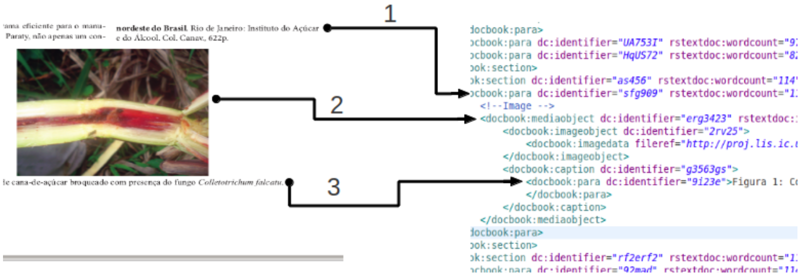
4. Previous Work

This proposal was conceived after experiences acquired during three previous research projects. Although with different strategies, they addressed complementary issues concerning data integration. In each project, experiments were conducted in a progressive integration fashion, starting from independent artifacts – represented by proprietary formats, in many cases – going towards the production of connections in lightweight or heavyweight integration approaches. As we will show here, our heavyweight integration here took a different perspective from an upfront one-step integration. It is the end of a chain of integration steps, in which the semantics inferred from the content in the first integration steps influences the following integration steps.

We further detail and discuss the role of each work in the LinkedScales architecture. While [Mota and Medeiros 2013] explores a homogeneous representation model for textual documents independently of their formats, [Bernardo et al. 2013] and [Miranda and Santanchè 2013] focus, respectively, on extracting and recognizing relevant information stored in spreadsheets and XML artifacts, to exploit their latent semantics in integration tasks.

4.1. Homogeneous Model – Universal Lens for Textual Document Formats

One of the key limits to index, handle, integrate and summarize sets of documents is the heterogeneity of their formats. In order to address this problem, we envisaged a "document space" in which several document sources represented in heterogeneous formats are mapped to a homogeneous model we call *Shadow* [Mota and Medeiros 2013].


1
2
3
ama eficiente para o manu- Paraty, não apenas um con-
nordeste do Brasil. Rio de Janeiro: Instituto do Açúcar e do Álcool. Col. Canav., 622p.
le cana-de-açúcar broqueado com presença do fungo Colletotrichum falcatu

```
docbook:para>
ocbook:para dc:identifier="UA753I" rstextdoc:wordcount="9
ocbook:para dc:identifier="HqUS72" rstextdoc:wordcount="8
ook:section>
ok:section dc:identifier="as456" rstextdoc:wordcount="114"
ocbook:para dc:identifier="sfg909" rstextdoc:wordcount="1
  <!--Image -->
  <docbook:mediaobject dc:identifier="erg3423" rstextdoc:
      <docbook:imageobject dc:identifier="2rv25">
          <docbook:imagedata fileref="http://proj.lis.ic.
      </docbook:imageobject>
      <docbook:caption dc:identifier="g3563gs">
          <docbook:para dc:identifier="9i23e">Figura 1: C
          </docbook:para>
      </docbook:caption>
  </docbook:mediaobject>
docbook:para>
ook:section>
ok:section dc:identifier="rf2erf2" rstextdoc:wordcount="1
ocbook:para dc:identifier="92mad" rstextdoc:wordcount="11
```

Figure 3. Main idea behind the work [Mota and Medeiros 2013]: A PDF document and its corresponding shadow.

Figure 3 illustrates a typical Shadow (serialized in XML). The content and structure of a document in a specific format (e.g., PDF, ODT, DOC) is extracted and mapped to an open structure – previously defined. The model behind this new structure, which is homogeneous across documents in the space, is a common hierarchical denominator found in most textual documents – e.g., sections, paragraphs, images. In the new document space a shadow represents *format+structure* of a document, decoupled from its specialized format.

Shadows documents are abstractions of documents in specific formats, i.e., they do not represent integrally the information of the source, focusing in the common information that can be extracted according to the context. This abstract homogeneous model allowed us to develop interesting applications in: document content integration and semantic enrichment [Mota et al. 2011]; and searching in a document collection considering structural elements, such as labels of images or references [Mota and Medeiros 2013].

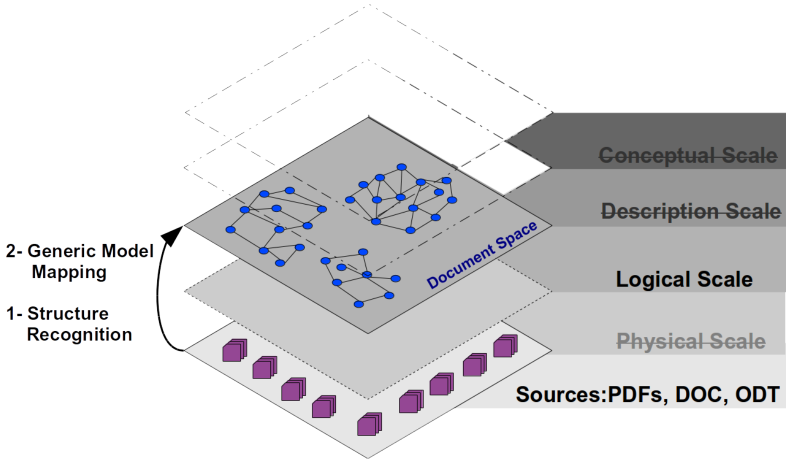


Figure 4. Shadows approach presented in a LinkedScales perspective.

Figure 4 illustrates how this homogeneous view for a document space fits in the LinkedScales architecture. This document space is equivalent to the Logical Scale, restricted to the document context. Different from the LinkedScales approach, Shadows map the documents in their original format straight to the generic model, without an intermediary Physical Scale.

After the Shadows experience we observed three important arguments to represent such intermediary scale: (i) since this scale is not aimed at mapping the resources to a common model, it focus in the specific concern of making explicit and addressable the content; (ii) it preserves the best-effort graph representation of the source, with provenance benefits; (iii) the big effort in the original one-batch-way conversion is factored in smaller steps with intermediary benefits.

In the LinkedScales' *Logical Scale*, the Shadows' document-driven common model will be expanded towards a generic perspective involving a family of models.

4.2. Connecting descriptive XML data – a Linked Biology perspective

[Miranda and Santanchè 2013] studied a particular problem in the biology domain, related to phenotypic descriptions and their relations with phylogenetic trees. Phenotypic descriptions are a fundamental starting point for several biology tasks, like identification of living beings or phylogenetic tree construction. Tools for this kind of description usually store data in independent files following open standards (e.g., XML). The descriptions are still based on textual sentences in natural language, limiting the support of machines in integration, correlation and comparison operations.

Even though modern phenotype description proposals are based on ontologies, there still are open problems of how to take advantage of the existing patrimony of descriptions. In such scenario, [Miranda and Santanchè 2013] proposes a progressive integration approach based on successive graph transformations, which exploits the existing latent semantics in the descriptions to guide this integration and semantic enrichment.

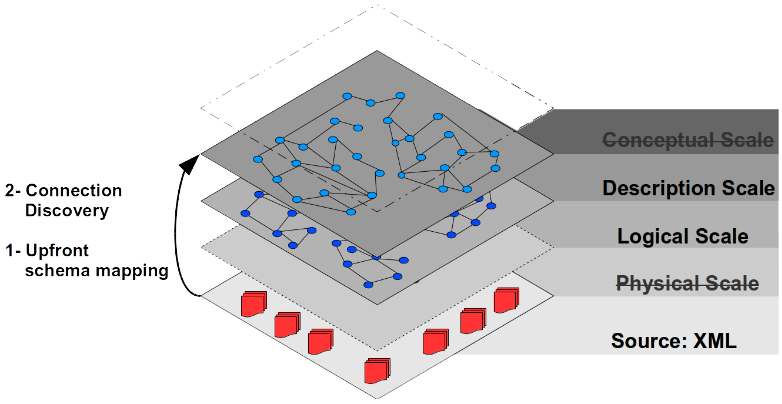


Figure 5. Linked Biology project presented in a LinkedScales perspective.

Since the focus is in the content, this approach departs from a graph-based schema which is a minimal common denominator among the main phenotypic description standards. Operations which analyses the content – discovering hidden relations – drive the integration process. Figure 5 draws the intersection between our architecture and the integration approach proposed by [Miranda and Santanchè 2013]. Data of the original artifacts are mapped straight to the Description Scale, in which structures have a secondary role and the focus is in the content.

In spite of the benefits of the focus in the content, simplifying the structures, this approach loses information which will be relevant for provenance. Moreover, in an interactive integration process, the user can perceive the importance of some information not previously considered in the Description Scale. In this case, since the mapping comes straight from the original sources, it becomes a hard task to update the extraction/mapping algorithms to afford each new requirement. The Physical and Logical Scales simplify this interactive process, since new requirements means updating graph transformations from lower to upper scales.

4.3. Progressively Integrating Biology Spreadsheet Data

Even though spreadsheets play important role as "popular databases", they were designed as self contained units. This characteristic becomes an obstacle when users need to integrate data from several spreadsheets, since the content is strongly coupled to file formats, and schemas are implicit driven to human consumption. In [Bernardo et al. 2013], we decoupled the content from the structure to discover and make explicit the implicit schema embedded in the spreadsheets.

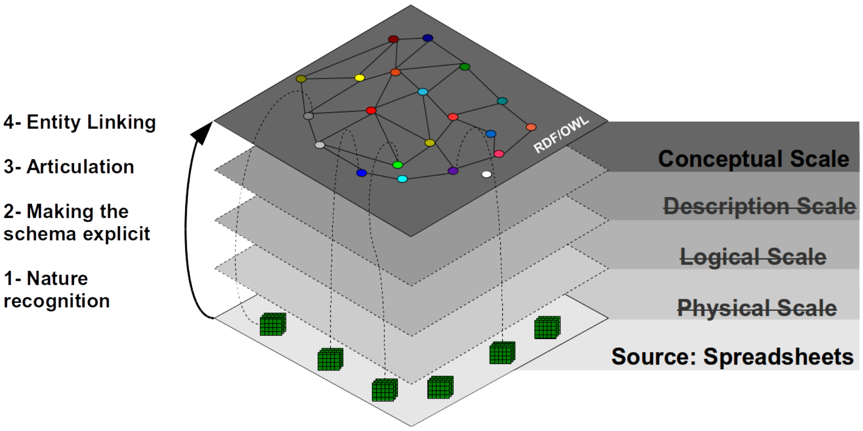
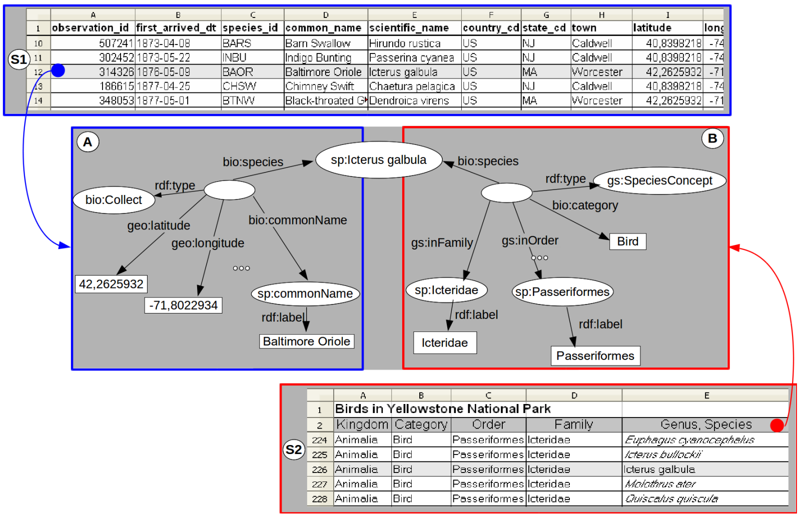


Figure 6. Spreadsheet integration presented in a LinkedScales perspective.

Figure 6 illustrates the [Bernardo et al. 2013] approach in a LinkedScales perspective. The work is divided in four steps, going from the original spreadsheets formats straight to the *Conceptual Scale*. The first step is to recognize the spreadsheet nature. The work assumes that users follow and share domain-specific practices when they are constructing spreadsheets, which result in patterns to build them. Such patterns are exploited in order to capture the nature of the spreadsheet and to infer a conceptual model behind the pattern, which will reflect in an ontology class in the *Conceptual Scale*.



	A	B	C	D	E	F	G	H	I	
1	observation_id	first_arrived_dt	species_id	common_name	scientific_name	country_cd	state_cd	town	latitude	long
10	507241	1873-04-08	BARS	Barn Swallow	Hirundo rustica	US	NJ	Caldwell	40,8398218	-74
11	302452	1873-05-22	INBU	Indigo Bunting	Passerina cyanea	US	NJ	Caldwell	40,8398218	-74
12	314326	1876-05-09	BAOR	Baltimore Oriole	Icterus galbula	US	MA	Worcester	42,2625932	-71
13	186615	1877-04-25	CHSW	Chimney Swift	Chaetura pelagica	US	NJ	Caldwell	40,8398218	-74
14	348053	1877-05-01	BTNW	Black-throated G	Dendroica virens	US	MA	Worcester	42,2625932	-71

	A	B	C	D	E
1	**Birds in Yellowstone National Park**				
2	Kingdom	Category	Order	Family	Genus, Species
224	Animalia	Bird	Passeriformes	Icteridae	*Euphagus cyanocephalus*
225	Animalia	Bird	Passeriformes	Icteridae	*Icterus bullockii*
226	Animalia	Bird	Passeriformes	Icteridae	Icterus galbula
227	Animalia	Bird	Passeriformes	Icteridae	*Molothrus ater*
228	Animalia	Bird	Passeriformes	Icteridae	*Quiscalus quiscula*

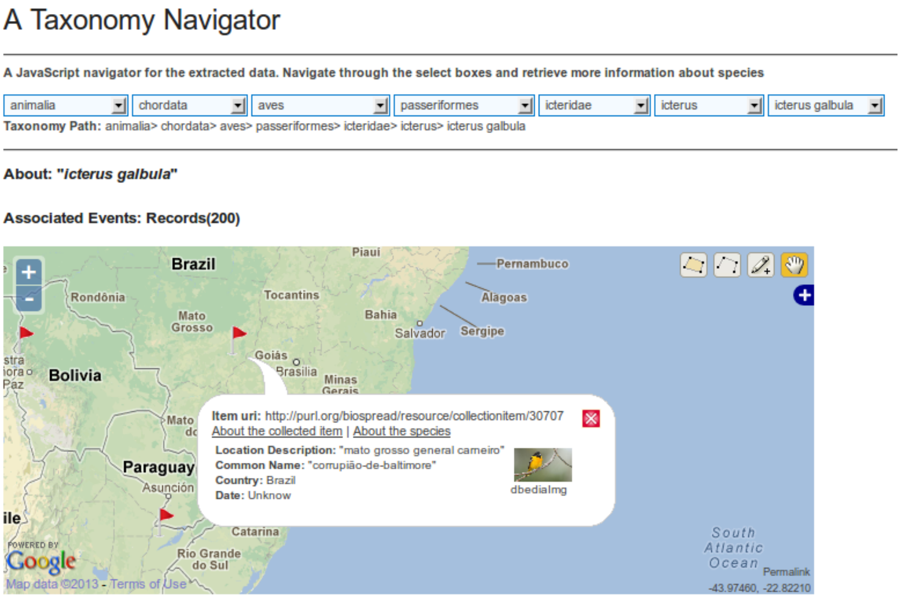
Figure 7. Spreadsheet data articulation via entity recognition.

This work stresses the importance of recognizing data as semantic entities to guide further operations of integration and articulation. Via this strategy, authors are able to transform several spreadsheets into a unified and integrated data repository. Figure 7 shows an example summarizing how they are articulated, starting from the recognition of semantic entities behind implicit schemas. Two different spreadsheets ($S1$ and $S2$) related to the biology domain have their schema recognized and mapped to specific ontology classes – shown in Figure 7 as (A) and (B).

Semantic entities can be properly interpreted, articulated and integrated with other sources – such as DBPedia, GeoSpecies and other open datasets. In an experiment involving more than 11,000 spreadsheets, we showed that it is possible to automatically recognize and merge entities extracted from several spreadsheets.

Figure 8 shows a screencopy of our query and visualization prototype for data[3] extracted from spreadsheets (available in `http://purl.org/biospread/?task=pages/txnavigator`).

This work subsidized our proposal of a Conceptual Scale as the topmost layer of our LinkedScales architecture. Several intermediary steps of transformation from the original datasources towards entities are hidden inside the extraction/mapping program. As in the previous cases, the process can be improved by materializing these intermediate steps in scales of our architecture.



A Taxonomy Navigator

A JavaScript navigator for the extracted data. Navigate through the select boxes and retrieve more information about species

animalia | chordata | aves | passeriformes | icteridae | icterus | icterus galbula

Taxonomy Path: animalia> chordata> aves> passeriformes> icteridae> icterus> icterus galbula

About: "*icterus galbula*"

Associated Events: Records(200)

Item uri: http://purl.org/biospread/resource/collectionitem/30707
About the collected item | About the species

Location Description: "mato grosso general carneiro"
Common Name: "corrupião-de-baltimore"
Country: Brazil
Date: Unknow

Figure 8. Screencopy of our prototype integrating data of several spreadsheets.

5. Concluding Remarks

This work presented a proposal for a dataspace system architecture based on graphs. It systematizes in layers (scales) progressive integration steps, based in graph transformations. The model is founded in previous work, which explored different aspects of the proposal. LinkedScales is aligned with the modern perspective of treating several heterogeneous datasources as parts of the same dataspace, addressing integration issues in progressive steps, triggered on demand. Although our focus is in the architectural aspects, we are designing a generic architecture able to be extended to several contexts.

Acknowledgments

Work partially financed[4] by CNPq (#141353/2015-5), 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.

[3] All data is available at our SPARQL endpoint: http://sparql.lis.ic.unicamp.br

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

References

[Angles 2012] Angles, R. (2012). A comparison of current graph database models. In *Data Engineering Workshops (ICDEW), 2012 IEEE 28th International Conference on*, pages 171–177. IEEE.

[Angles and Gutierrez 2008] Angles, R. and Gutierrez, C. (2008). Survey of graph database models. *ACM Comput. Surv.*, 40(1):1:1–1:39.

[Belhajjame et al. 2013] Belhajjame, K., Paton, N. W., Embury, S. M., Fernandes, A. A., and Hedeler, C. (2013). Incrementally improving dataspaces based on user feedback. *Information Systems*, 38(5):656 – 687.

[Bernardo et al. 2013] 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 Information and Data Management*, 4(2):104.

[Cai et al. 2005] Cai, Y., Dong, X. L., Halevy, A., Liu, J. M., and Madhavan, J. (2005). Personal information management with semex. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, pages 921–923, New York, NY, USA. ACM.

[Dijk et al. 2013] Dijk, J., Choenni, S., Leertouwer, E., Spruit, M., and Brinkkemper, S. (2013). A data space system for the criminal justice chain. In Meersman, R., Panetto, H., Dillon, T., Eder, J., Bellahsene, Z., Ritter, N., Leenheer, P., and Dou, D., editors, *On the Move to Meaningful Internet Systems: OTM 2013 Conferences*, volume 8185 of *Lecture Notes in Computer Science*, pages 755–763. Springer Berlin Heidelberg.

[Dittrich et al. 2009] Dittrich, J., Salles, M. A. V., and Blunschi, L. (2009). imemex: From search to information integration and back. *IEEE Data Eng. Bull.*, 32(2):28–35.

[Dong and Halevy 2007] Dong, X. and Halevy, A. (2007). Indexing dataspaces. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, SIGMOD '07, pages 43–54, New York, NY, USA. ACM.

[Elsayed and Brezany 2010] 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.

[Elsayed et al. 2006] Elsayed, I., Brezany, P., and Tjoa, A. (2006). Towards realization of dataspaces. In *Database and Expert Systems Applications, 2006. DEXA '06. 17th International Workshop on*, pages 266–272.

[Franklin et al. 2005] Franklin, M., Halevy, A., and Maier, D. (2005). From databases to dataspaces. *ACM Sigmod Record*, 34(4).

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

[Halevy et al. 2006b] Halevy, A., Rajaraman, A., and Ordille, J. (2006b). Data integration: The teenage years. In *Proceedings of the 32Nd International Conference on Very Large Data Bases*, VLDB '06, pages 9–16. VLDB Endowment.

[Heath and Bizer 2011] Heath, T. and Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*.

[Hedeler et al. 2009] Hedeler, C., Belhajjame, K., Fernandes, A., Embury, S., and Paton, N. (2009). Dimensions of dataspaces. In Sexton, A., editor, *Dataspace: The Final Frontier*, volume 5588 of *Lecture Notes in Computer Science*, pages 55–66. Springer Berlin Heidelberg.

[Hedeler et al. 2010] Hedeler, C., Belhajjame, K., Paton, N., Campi, A., Fernandes, A., and Embury, S. (2010). Chapter 7: Dataspaces. In Ceri, S. and Brambilla, M., editors, *Search Computing*, volume 5950 of *Lecture Notes in Computer Science*, pages 114–134. Springer Berlin Heidelberg.

[Hedeler et al. 2013] Hedeler, C., Fernandes, A., Belhajjame, K., Mao, L., Guo, C., Paton, N., and Embury, S. (2013). A functional model for dataspace management systems. In Catania, B. and Jain, L. C., editors, *Advanced Query Processing*, volume 36 of *Intelligent Systems Reference Library*, pages 305–341. Springer Berlin Heidelberg.

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

[Kolaitis 2005] Kolaitis, P. G. (2005). Schema mappings, data exchange, and metadata management. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, PODS '05, pages 61–75, New York, NY, USA. ACM.

[Lenzerini 2002] Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '02, pages 233–246, New York, NY, USA. ACM.

[Madhavan et al. 2007] Madhavan, J., Cohen, S., Dong, X. L., Halevy, A. Y., Jeffery, S. R., Ko, D., and Yu, C. (2007). Web-scale data integration: You can afford to pay as you go. In *CIDR*, pages 342–350. www.cidrdb.org.

[Miranda and Santanchè 2013] Miranda, E. and Santanchè, A. (2013). Unifying phenotypes to support semantic descriptions. *Brazilian Conference on Ontological Research – ONTOBRAS*, pages 1–12.

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

[Mota et al. 2011] Mota, M. S., Longo, J. S. C., Cugler, D. C., and Medeiros, C. B. (2011). Using linked data to extract geo-knowledge. In *GeoInfo*, pages 111–116.

[Singh and Jain 2011] 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.