Progressive Data Integration and Semantic Enrichment Based on LinkedScales and Trails

Matheus Silva Mota, Fagner Leal Pantoja,
Julio Cesar dos Reis, and André Santanchè

Institute of Computing, University of Campinas, São Paulo, Brazil
{mota;pantoja;julio.dosreis;santanche}@ic.unicamp.br

Abstract. The integration of data elements scattered along different resources, with heterogeneous formats, can take advantage of an approach with progressive and lightweight steps, instead of pursuing costly upfront mappings. To support such approach, we defined a multiscale-based dataspace architecture, called *LinkedScales*, which carries an integration process via graph-based transformations over a graph database. A series of scales in the dataspace systematizes an integration and enrichment chain of steps to leverage transformation processes, which incrementally go from raw representations towards ontology-like structures. However, how to record and keep track of the intermediary outcomes in the integration chain remains an open research challenge. This article proposes combining the concept of scales with trails – lightweight, scale-specialized semantic annotations to enable progressive integration towards a semantic representation. We conduct experiments involving organism-centric analysis in life science to show the benefits of trails for transformation between scales.

Keywords: Data Integration, Dataspaces, Multiscale, Organism-centric analysis, Trails, Semantic annotation

1 Introduction

Biologists often conduct organism-centric analysis in which organisms – *i.e.*, species or taxonomic groups – are the central focus and data are collected and integrated around them. In this context, biologists might compare organisms in a systematic way and investigate conditions related to their hypotheses. In this context, the construction of *profiles* [10] as "*views*" of data is usual in an organism-centric research. It involves combining data usually fragmented in heterogeneous sources, requiring efforts to collect and combine pieces coming from multiple repositories and files with different formats. The manual process requires a lot of time to prepare data from each source and to integrate them before any analysis. Fig. 1 presents a practical scenario where the analysis is based on profiles comprising ecological traits and morphological data. It requires the combination of data from several resources scattered in digital repositories. In this case, the data comes from research repositories associated with scientific publications, such as *Dryad* (http://datadryad.org) and *Figshare* (http://figshare.com).

The combination of datasets is challenging since the different kinds of heterogeneity, *i.e.*, distinct formats (CSV, Excel, NeXML), structures (tables, trees) and schemas, *etc.* require several steps of integration.

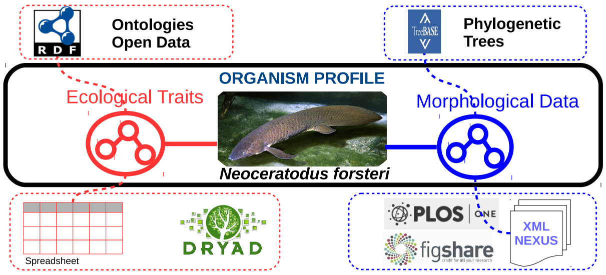


Fig. 1. Profile integrating characteristics scattered across several sources

Heterogeneity hampers a unified exploration of knowledge across distinct systems. To provide an on-demand lightweight integration, we have defined the *LinkedScales* architecture [4], which aims at splitting the integration steps as discrete scales. Each scale encompasses common aspects and routines related to a particular integration step. *LinkedScales* comply with the dynamicity of modern integration environments, against the classic heavyweight upfront techniques.

This incremental process also produces three kinds of intermediary outcomes: semantic representations, knowledge discovery results and user feedback. They have operational purposes and drive transformation tasks in the production of content in the upper scales. However, there is no a systematic method to record and keep track of these intermediary outcomes. Operations built over them, like transformation and enrichment, can be better specified, managed and followed if they rely on a standard mechanism to document the outcomes.

In this article, we propose combining *LinkedScales* with the concept of trails. Trails are "*hints*" represented as structured semantic annotations concerning operational scale aspects – *i.e.*, each scale emphasizes a particular step of the integration chain, therefore each scale has distinct types of trails. Trails play the role of metadata associated with portions of data [1]. When trails are included in a progressive integration process, they standardize the way in which intermediary results are represented, which might improve the specification of transformation rules. Furthermore, *LinkedScales* produces a provenance graph while transformations are executed. This graph contains not only information about processes, but also which operational evidence (trails) were considered during the transformation.

We present a practical scenario of exploring trails with *LinkedScales*. We conducted an experimental analysis considering the integration and semantic enrichment of resources related to a particular organism profile. In particular, trails are exploited to guide the process of linking content in the scales with external knowledge bases, like *DBpedia*, to better characterize the data conceptually. In order to show how trails can improve the linking process, in the first step of our experimental procedure, we apply the transformation without the trails and we compare the results taking the trails into account afterward.

The remaining of this article is organized as follows: Section 2 presents foundations and related work. Section 3 describes the *LinkedScales* framework while Section 4 details the proposal of combining *LinkedScales* with trails. Section 5 presents the conclusion remarks.

2 Foundations and Related Work

Several data integration approaches have emerged, including federated databases, schema integration and data warehouses [7]. They mostly rely on providing a virtual unified view under a global schema (GS) [8]. Within GS-based systems, data stay in their original data sources – *i.e.*, maintaining their original schemas – and are dynamically fetched and mapped to a global schema [2]. It requires a big upfront effort to produce a global schema definition, which may become impracticable due to the inclusion and changes in schemas. Such classical data integration might successfully work when integrating modest numbers of stable databases in controlled environments.

Scenarios in which schemas often change and new data models must be considered still lack an efficient solution. To this end, pay-as-you-go integration approaches implement incremental integration based on progressive steps to continuously refine and improve the connections among sources. The proposal of *dataspaces* aims at providing the benefits of the classical data integration approach but in a progressive fashion way [8]. Dataspaces approach for data integration can be divided into a bootstrapping stage and subsequent refinements. Progressive integration refinements may rely on structural analysis, on user feedback or on manual/automatic mappings among sources [1].

This investigation explores the concept of trails in a pay-as-you-go integration approach. Trails are keyword-based annotations that relate concepts to data sources to be integrated. They are used for a gradual improvement of integration among sources [9]. Trails play a key role since an important step in integration tasks involves defining semantic equivalences across distinct data sources during the dataspace improvement. In some proposals, the user is engaged in helping the semi-supervised process of discovering, suggesting and evaluating mappings, either by statistical techniques or driven by ontologies and dictionaries [1].

As an alternative for the one-step approach to define equivalences between distinct data source elements, trails rely on *services* to support incremental refinements of mappings between schemas. Whenever the user feeds the system with new "hints", it exploits them to improve the semantic equivalences discovery. These "hints" are treated as a lightweight mechanism to define declarative relationships between loosely integrated data sources [1]. Trails can be associated with either a particular portion of the data or the whole dataset. They can be either automatically inferred or manually assigned, depending on the effort that users are willing to spend [9].

3 LinkedScales

LinkedScales [4] refers to an architecture that systematizes the progressive integration steps, bringing the proposal of multiscale to the data integration chain. It is based on an abstract model that organizes the integration steps as a pile of scales, where the entities in an upper scale are built based on transformations over entities of a lower scale.

The integration starts on the lowest scale, where all original data sources are ingested and transformed into graphs. Each subsequent scale from this point is a graph derived from the previous scale, taking advantage of the flexibility of graphs to logically represent different structures along the scales. This model allows representing operations within and across the scales as transformation procedures in graphs. Fig. 2 presents the four scales aiming at going from the raw data sources (lower scales, containing more details about format and structure) to a conceptual scale (fewer details of format and structure, and focus on domain-specific concepts). Scales are interconnected by an orthogonal graph, supporting traceability among them – *i.e.*, it is possible to "track" sources/targets of transformations between scales.

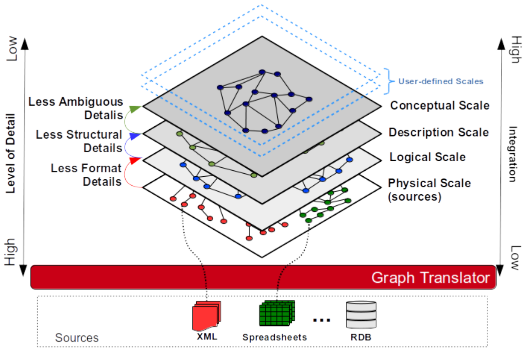


Fig. 2. *LinkedScales Primary Data Architecture* [4]

The ***Physical Scale*** aims at representing the different data sources in their original physical format as a graph. The original raw data sources are transformed into a graph by an ingestion procedure. The Graph Translator reads several specialized formats – *e.g.*, Excel, CSV, relational tables, XML – and converts them to an equivalent graph representation. The original structure, format and content of the underlying data sources are reflected in a graph.

The ***Logical Scale*** offers a common view for data inside similar or equivalent logical models represented in the previous scale. Tables and hierarchical documents are examples of logical models present in the sources. 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 an HTML file since they preserve specificities of their formats.

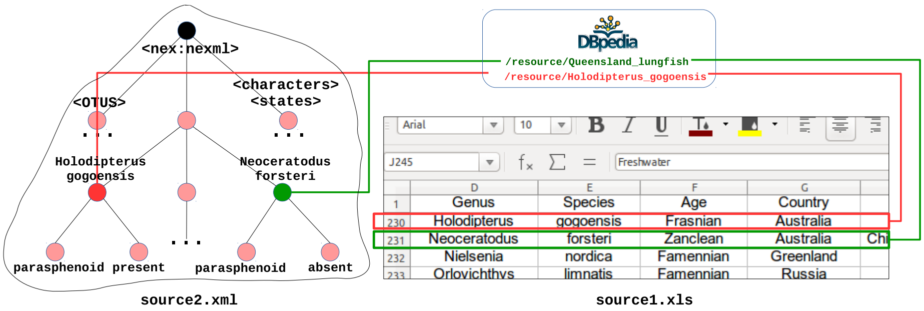
The ***Description Scale*** emphasizes the content (*e.g.*, labels of elements within an XML document or values in spreadsheet cells) and their relationships. Since models represent relations among data elements in different ways – *e.g.*, a row in a table can represent data concerning the same entity while hierarchical relations in a document represent aggregations – the *Description Scale* reduces all logical models to a single unified one, to shift the focus towards the descriptive content. The unified model selected for this scale relies on the triple <resource, property, value>, which is usual in several meta-data standards as *Resource Description Framework* (RDF).

The highest scale refers to the ***Conceptual Scale***. It integrates data from the lower scale at a semantic level by exploiting the content and relationships between nodes to discover and make explicit the semantics through ontologies. Entities are discovered, deduplicated and related to ontologies as instances of classes, or properties and their values. A "textual graph" of the previous scale becomes a graph containing interrelated entities and their properties/values, with explicit semantics supported by ontologies.

4 Combining Trails with LinkedScales

This work involves an enhancement of the *LinkedScales* framework to incorporate *Trails* as the driving component for transformations and provenance. It treats trails as scale-specialized operational semantic annotations, which indicates the role of data portions. Such *hints* are considered by scale transformation processes, incrementally conducting the refinement of the dataspace.

We conducted an experiment in the organism-centric scenario to investigate how trails improve transformations between scales. We collected two complementary sources coming from different scientific publications – as illustrated in Figure 3. The first source is an XLS spreadsheet [3] shared in the Dryad repository. The second source is a NeXML file – an XML-based format for representing phylogenetic and phenotypic data, shared in the Figshare repository [6].



	D	E	F	G	
1	Genus	Species	Age	Country	
230	Holodipterus	gogoensis	Frasnian	Australia	
231	Neoceratodus	forsteri	Zanclean	Australia	Chi
232	Nielsenia	nordica	Famennian	Greenland	
233	Orlovichthys	limnatis	Famennian	Russia	

Fig. 3. Schematic illustration of the bootstrap phase of the experiment

Both data sources are concerned with information about lungfish. While the first data source contains morphological traits, behavioral aspects, habitat characteristics, *etc.* of several lungfish species, the second data source comprises a phylogenetic tree and a phenotypic description in a character/character state format. Even though both data sources are available for researchers, integrating such information conceptually by combining data of the same lungfish species remains a challenging laborious task.

In next sections we exploit the data sources as a running example to describe the use of different types of trails and their relationship with scales. Trails vary according to each scale, indicating relevant aspects of data that the transformation process takes into account during the production of an upper scale. We further describe roles of trails presenting the scale that they are inserted accompanied by the target transformation scale –*e.g.*, a *physical-logical trail* refers to a trail to be inserted in the physical scale, impacting in the data production of the logical scale.

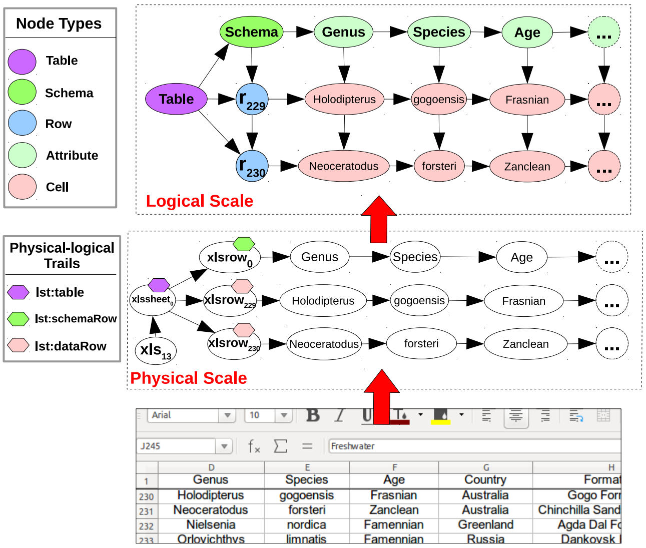
4.1 Physical-logical Trails

Lowest part of Figure 4 presents an excerpt of an XLS spreadsheet containing information from a study of discrete characters change in the evolution of lungfish (class *Sarcopterygii*) [3]. The dataset is an asset associated with a publication, shared in the *Dryad* repository. It describes information about taxonomic classification, associated geological age, type of habitat, countries, *etc.*

Data is ingested into *LinkedScales* database as a graph. The middle part of Figure 4 shows partial representation of the ingestion result in the *Physical Scale.* Rows of nodes represent rows of the spreadsheet and their stream of cells. The graph focus on representing as much information as possible of the raw resource. Via such data, the logical organization can be inferred or derived – *e.g.*, initial and boldly formatted cells usually are the table schema.

Physical-logical trails – pictured as colored hexagons in Figure 4 – are inserted to distinguish types of structures and their internal components. Figure 4 (middle part) illustrates how trails are used to conduct transformations from the physical to the logical scale. Trails associate structure-related roles to the nodes as: table (*lst:table*), row (*lst:dataRow*) and the stream of cells corresponding to the schema (*lst:schemaRow*). In the bootstrap phase of the dataspace, this type of trail is either automatically inferred by the ingestion module, according to the internal structures, or specified by the user. In short, the Physical-logical trails indicate how data is logically organized within the format-specific graph representation of the resource.

Based on the associated physical-logical trails, a transformation process adopts a standard representation of structural elements of *tables* to logically represent the resource in the logical scale. Representing structures using a standard representation in the logical scale is particularly important, as it allows, for instance, reusing table-related algorithms to reach resources independently of formats.



	D	E	F	G	H
1	Genus	Species	Age	Country	Format
230	Holodipterus	gogoensis	Frasnian	Australia	Gogo Forr
231	Neoceratodus	forsteri	Zanclean	Australia	Chinchilla Sand
232	Nielsenia	nordica	Famennian	Greenland	Agda Dal Fo
233	Orlovichthys	limnatis	Famennian	Russia	Dankovsk I

Fig. 4. Excerpt of a XLS and its representation on the *Physical* and *Logical scales*

4.2 Logical-description Trails

The *Description Scale* aims at shifting the focus to the content and their relationships, reducing logical models to an RDF-based structure. The bottom part of Figure 5 illustrates how **logical-description trails** are used to produce the description scale from the logical scale. At this point, trails indicate how structural elements should be organized as <*resource, property, value*> triples.

Figure 5 illustrates how trails (colored hexagons) are associated with structural elements on the *Logical Scale*, indicating, for instance, that rows (nodes r_{229} and r_{230}) are resources, schema attributes (green nodes *Genus*, *Species*, *Age*) are properties and cells are values – *e.g.,* $< r_{230}, Genus, Neoceratodus >$ and $< r_{230}, Species, forsteri >$ are triples produced based on trails.

The transformation illustrated in Figure 5 can be represented by a rule which matches a pattern (including specific trails) as input and produces a transformed output. Transformation patterns are already defined in the *LinkedScales* model [4], and are beyond the scope of this work.

4.3 Description-conceptual Trails

Description-conceptual trails focus on reaching an expected perspective – e.g., organism profiles. Figure 5 (upper part) illustrates trails indicating the expected semantic interpretation of nodes in the description scale, making the semantic explicit by adopting specific elements of ontologies. Such trails can be automatically discovered by the system in a semi-supervised process or be directly assigned.

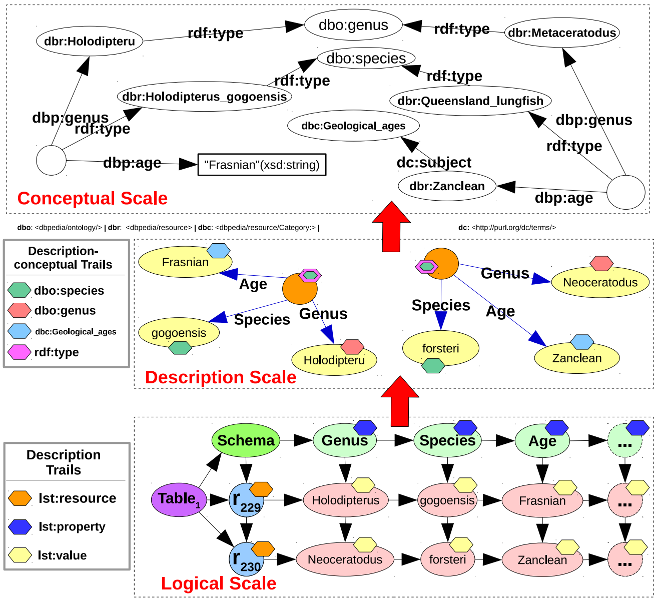


Fig. 5. Logical-description trails driving a logical-description scale transformation, and description-conceptual trails driving the production of the conceptual scale

The *Conceptual Scale* addresses fundamental semantic concerns by distinguishing entities and adopting controlled vocabularies to represent descriptive properties. Adjustments – removing or adding description-conceptual trails – made on previous scale are a way for handling the dynamicity of scenarios as organims-centric research in terms of testing different hypothesis.

Scales and trails play complementary roles in the progressive integration process. While a scale provides a homogeneous view of the lower layers, trails offer the proper clues for the transformation to the next scale. Consider, for example, the logical scale. It offers a homogeneous view of data considering the logical model, i.e., all tables are represented in the same way, as well as, all trees. If on one hand, this is a powerful mechanism, as the heterogeneity of several table formats is hidden in a lower scale, enabling to reuse the same algorithms for several homogeneous tables, on the other hand, these algorithms need clues to interpret implicit differences which will impact in the next scale.

Regarding the experiment of integrating both XLS and NeXML sources, at the bootstrap stage, after ingesting both data files and converting them to the Logical and Description scales in the graph, we used *DBpedia* (dbpedia.org) to automatically produce the trails that guided the production of the Conceptual scale. The experiment aimed at connecting portions of the data source with DBPedia resources (English release of October 2015), and therefore indirectly linking and enriching similar resources.

Our procedure searches in the *DBpedia* for the most similar resources of each node in the graph. The search method compares the input query against the *DBpedia* resource contents. This comparison uses the *tf-idf* measure and may return approximate/incorrect results like uncorrelated resources. To examine the benefits brought by the trails in this transformation process, the next integration stage inserted trails associated to the nodes to give clues to our integration system about the nature of the nodes in the graph. In this experiment, two trails were considered associated with specific procedures:

-Species related Trail: The user tags the nodes that represent *Species*, then the system can filter, via SPARQL queries, the resources returned by the bootstrap stage that are instances of taxonomy-related classes, according to the *DBpedia* ontology.

-Morphological related Trail: The user tags the nodes that represent morphological characters. Such trails are used as input in an entity-quality recognition algorithm [5] that extracts morphological characters inside a free-text and creates an Entity-Quality (EQ) representation. The Entity element refers to the morphological character (*e.g.*, *bone*) and the Quality stands for a qualifier (*e.g.*, *present*) that specifies a given state of the Entity. The algorithm uses two domain-ontologies to support its recognition task: (1) Teleost Anatomy Ontology (TAO) to recognize the Entities and (2) Phenotypic Trait Ontology (PATO) to recognize the Qualities.

Figure 6 depicts a portion of the conceptual scale with (right part) and without (left part) trails. Each node in the figure represents a specific species from both data sources – the first data source in green, and the second data source in red – and edges represent relationships concerning taxonomy and morphological traits (entity-quality pairs). When trails are associated to elements of the previous scale (description), the produced conceptual scale is semantically refined according to the expected requirements in the organism profiles.

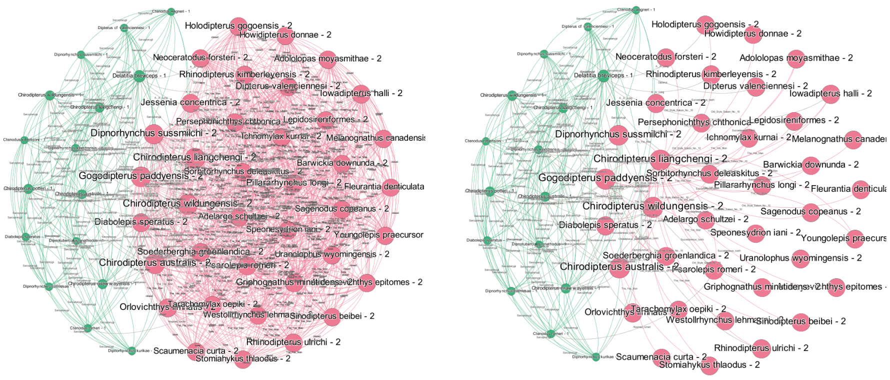

Holodipterus gogoensis - 2
Howidipterus donnae - 2
Neoceratodus forsteri - 2
Adololopas moyasmithae - 2
Rhinodipterus kimberleyensis - 2
Dipterus valenciennesi - 2
Iowadipterus halli - 2
Jessenia concentrica - 2
Persephonichthys chthonica - 2
Lepidosireniformes - 2
Dipnorhynchus sussmilchi - 2
Ichnomylax kurnai - 2
Melanognathus canadensis
Chirodipterus liangchengi - 2
Barwickia downunda - 2
Gogodipterus paddyensis - 2
Sorbitorhynchus deleaskitus - 2
Pillararhynchus longi - 2
Fleurantia denticulata
Chirodipterus wildungensis - 2
Sagenodus copeanus - 2
Adelargo schultzei - 2
Diabolepis speratus - 2
Speonesydrion iani - 2
Youngolepis praecursor
Soederberghia groenlandica - 2
Uranolophus wyomingensis - 2
Chirodipterus australis - 2
Psarolepis romeri - 2
Griphognathus minutidens - 2
Nielsenichthys epitomes - 2
Orlovichthys limnatis - 2
Tarachomylax oepiki - 2
Westollrhynchus lehmanni - 2
Sinodipterus beibei - 2
Rhinodipterus ulrichi - 2
Scaumenacia curta - 2
Stomiahykus thlaodus - 2
Ctenodus wagneri - 1
Dipterus cf. valenciennesi - 1
Dipnorhynchus sussmilchi - 1
Delatitia breviceps - 1
Chirodipterus wildungensis - 1
Chirodipterus liangchengi - 1
Ctenodus robertsoni - 1
Dipnorhynchus kiandrensis - 1
Chirodipterus potteri - 1
Chirodipterus australis - 1
Diabolepis speratus - 1
Dipnotuberculus gnathodus - 1
Dipnorhynchus cathlesae - 1
Chirodipterus onawayensis - 1
Ctenodus romeri - 1
Dipnorhynchus kurikae - 1

Fig. 6. Comparison of the same portion of two conceptual scales: Non trail-based (left) and trail-based (right) transformations

5 Conclusion

Asignificant part of the biological research work remains in an organism-centric perspective, which usually requires combining data regarding distinct aspects of organisms. In this article, we presented how our *LinkedScales* framework, based on the multiscale integration approach, can work aligned with trails as operational semantic annotations. Trails systematize intermediary outcomes, improving the transformation process and provenance records among the scales. Our experimental analysis demonstrated the overall potential benefits of trails in *LinkedScales* to reach organism profiles. Future work involves conducting additional experimental evaluations to thoroughly examine the quality and scalability of data integration provided by the approach.

Acknowledgments
Work partially financed by FAPESP (#2014/14890-0), INCT in Web Science, CNPq (#141353/2015-5), FAPESP-PRONEX (eScience project).

References

1. Belhajjame, K., Paton, N.W., Embury, S.M., Fernandes, A.A., Hedeler, C.: Incrementally improving dataspaces based on user feedback. Information Systems 38(5), 656 – 687 (2013)
2. Hedeler, C., Fernandes, A., Belhajjame, K., Mao, L., Guo, C., Paton, N., Embury, S.: A functional model for dataspace management systems. In: Catania, B., Jain, L.C. (eds.) Advanced Query Processing, Intelligent Systems Reference Library, vol. 36, pp. 305–341. Springer Berlin Heidelberg (2013)
3. Lloyd, G.T., Wang, S.C., Brusatte, S.L.: Identifying heterogeneity in rates of morphological evolution: Discrete character change in the evolution of lungfish (sarcopterygii; dipnoi). Evolution 66(2), 330–348 (2012)
4. Mota, M.S., dos Reis, J.C., Goutte, S., Santanchè, A.: Multiscaling a graph-based dataspace [accepted]. Journal of Information and Data Management - JIDM p. 16 (2016), `http://www.lis.ic.unicamp.br/wp-content/uploads/2016/10/multiscaling-graph-based.pdf`
5. Pantoja, F.L., Cavoto, P., Reis, J., Santanchè, A.: Generating Knowledge Networks from Phenotypic Descriptions. Proc. 12th IEEE e-Science pp. 1–10 (2016)
6. Pardo, J.D., Huttenlocker, A.K., Small, B.J.: An exceptionally preserved transitional lungfish from the lower permian of nebraska, usa, and the origin of modern lungfishes. PLoS ONE 9(9), 1–13 (09 2014)
7. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. The VLDB Journal 10(4), 334–350 (2001)
8. Singh, M., Jain, S.: A survey on dataspace. In: Wyld, D., Wozniak, M., Chaki, N., Meghanathan, N., Nagamalai, D. (eds.) Advances in Network Security and Applications, Communications in Computer and Information Science, vol. 196, pp. 608–621. Springer Berlin Heidelberg (2011)
9. Vaz Salles, M.A., Dittrich, J.P., Karakashian, S.K., Girard, O.R., Blunschi, L.: itrails: Pay-as-you-go information integration in dataspaces. In: Proceedings of the 33rd International Conference on Very Large Data Bases. pp. 663–674. VLDB '07, VLDB Endowment (2007)
10. Washington, N.L., Haendel, M.A., Mungall, C.J., Ashburner, M., Westerfield, M., Lewis, S.E.: Linking human diseases to animal models using ontologybased phenotype annotation. PLoS biology 7(11), e1000247 (2009)