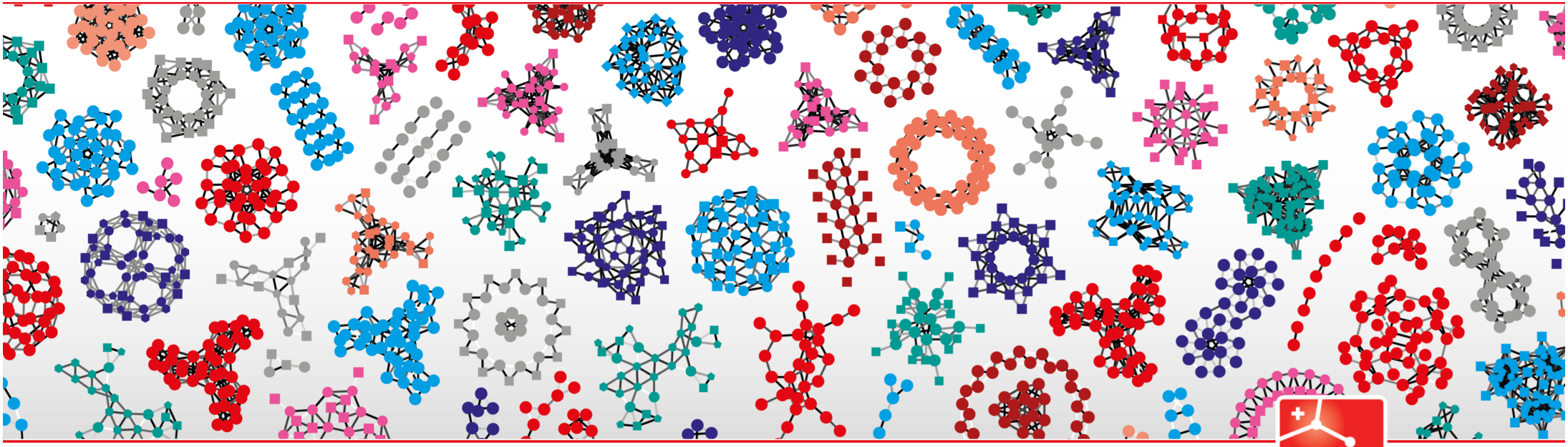


SIB Swiss Institute of Bioinformatics

Data integration in life-sciences: the current state

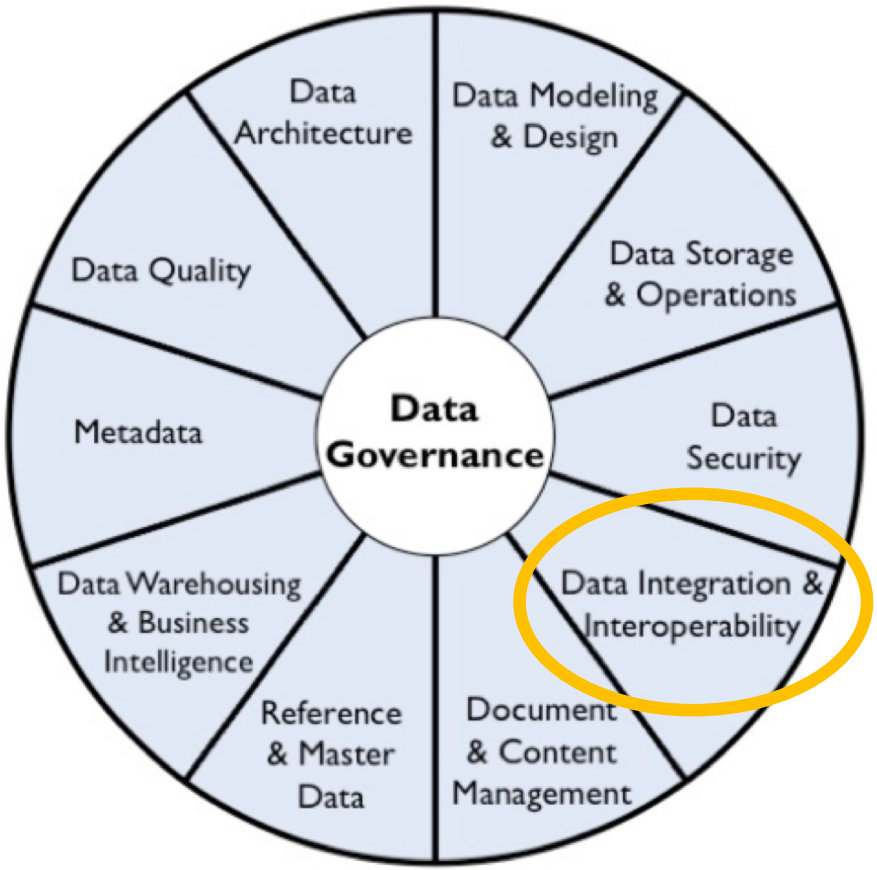
Tarcisio Mendes de Farias

tarcisio.mendes@sib.swiss

elixir SWITZERLAND | Unil UNIL | Université de Lausanne

Data integration and interoperability (DII)

- What is DII?
- Why DII is important?
  - Health care
  - Biology



What is DII?

- Data integration and Interoperability (DII)

  "describes processes related to the **movement and consolidation of data** within and between data stores, applications and organizations."[1]

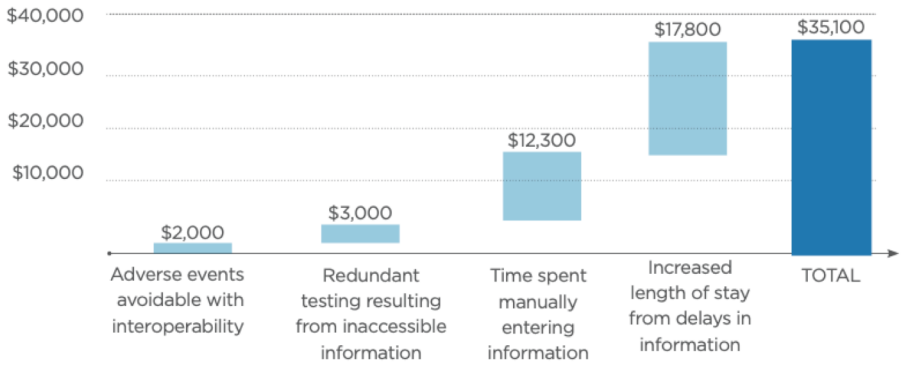
- **Data integration**: "consolidates data into consistent forms, either physical or virtual" [1]

- **Data Interoperability**: it "is the ability for multiple systems to communicate."[1]

[1]Henderson, D., Earley, S. & Sebastian-Coleman, L., 2017. *DAMA-DMBOK: Data management body of knowledge*, Technics Publications.

Why DII is important in health-care?

Figure 2: Estimated Addressable Waste
Estimated Waste from Lack of Medical Device Interoperability ($M)

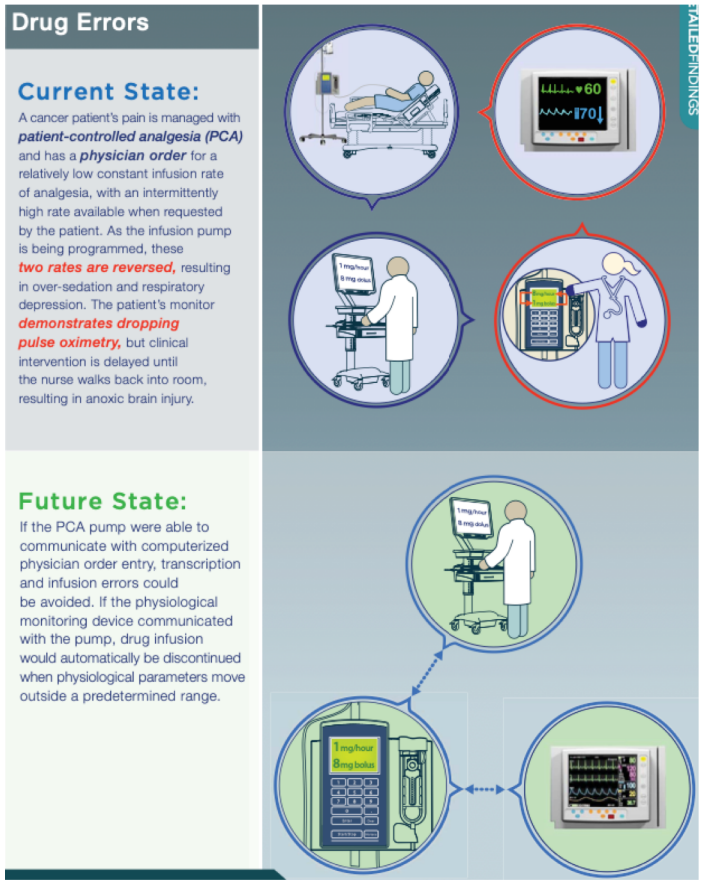


Estimated Waste from Lack of Commonly Adopted Standards ($M)

*Note: Numbers rounded for clarity

[1]https://www.westhealth.org/wp-content/uploads/2015/02/The-Value-of-Medical-Device-Interoperability.pdf

Providing means to personalised health-care

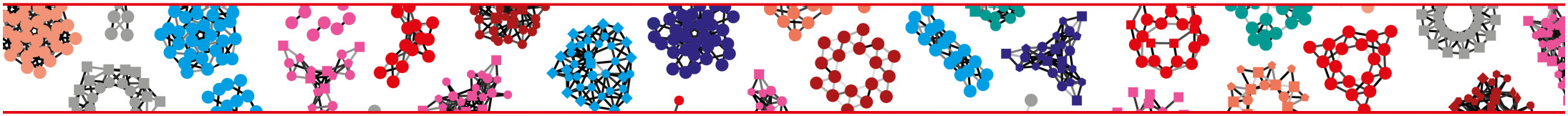


Why DII is important in biology?

- Considerably **reducing the fastidious and time-consuming tasks** of source discovery, to gather and **to combine data from different data sources**

- Enabling **knowledge discovery** and to answer **data-driven research** questions

- Promoting a widely **formalisation** and a consensus **of biological knowledge** through data standards



Challenges in *Biological* Data Integration

What are the main obstacles to achieve DII in biology?

- **HIGH** heterogeneity
  - context-specific requirements (i.e. no "one model fits all"), different data modelling decisions, domain-specific purposes, and technical constraints.

- Accessing a set of autonomous and heterogeneous data sources
  - e.g. challenge: offering a uniform way to query data

- Several dispersed biological datasets
  - Produced by different and autonomous research groups
  - E.g.:~50 databases about orthology; ~5 databases aggregating gene expression data around the globe.

Heterogeneity and Number of sources

- Database models, CSV files, **FASTQ** files, spreadsheets, HTML (webpages)
- Structured (1), semi-structured (2) and unstructured data (3)
  - E.g. different data schemas - (1) and (2)
  - E.g. full text - (2) and (3)
- Multitude of data sources on Web-scale
- **Semantic heterogeneity** – different data meanings and granularities

Data (value) heterogeneity

- Different ways to represent a data value

Big Blue

International Business Machine Corporation

Semantic heterogeneity

Fruit or vegetable?

- Keeping data sources autonomy and context (as they are), but still being interoperable
- **Utopic:** “One model fits all” approach

Semantic heterogeneity

- Different ways to structure a body of data, and consequently, data meanings
- Ambiguities

First Name	Middle Name	Family name
Tarcisio	null	Mendes de Farias

Name
Tarcisio Mendes de Farias

First Name	Mother's family name	Father's family name
Tarcisio	Mendes	de Farias

Heterogeneous data formats

- There are a multitude of sequence formats:
  - gcg, nbrf, pdb, swissprot, embl, fastq, gifast, refseqp, fastq-illumina, fastq-illumina…

FASTQ file - Illumina software

```
@HWUSI-EAS100R:6:73:941:1973#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATA
+HWUSI-EAS100R:6:73:941:1973#0/1
!"*((((***+))%%%++)(%%%%).1***-
```

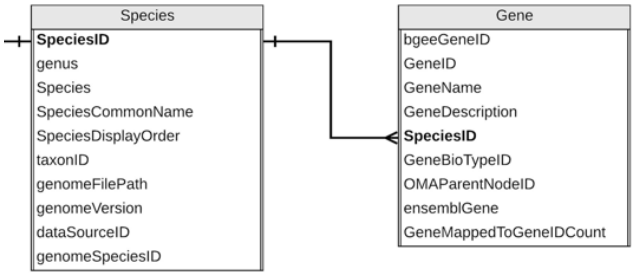
FASTQ file - INSDC Sequence Read Archive

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

the unique instrument name : flowcell lane : tile number within the flowcell lane : …
#0 -> index number for a multiplexed sample (0 for no indexing)
/1 ->the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

Heterogeneous database models

Relational model



MySQL

ORACLE DATABASE

PostgreSQL

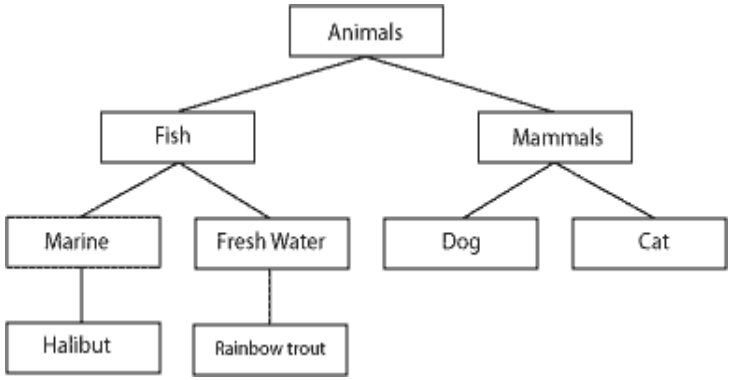
Graph model

blazegraph

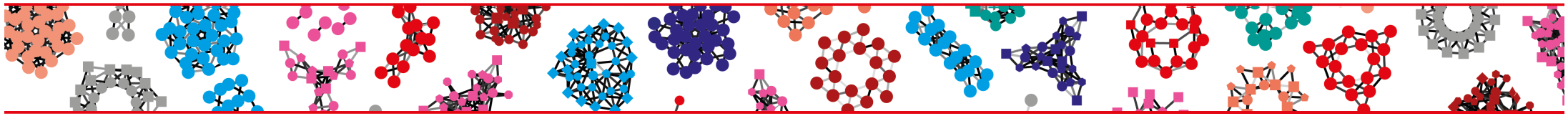
Stardog

neo4j

Hierarchical model



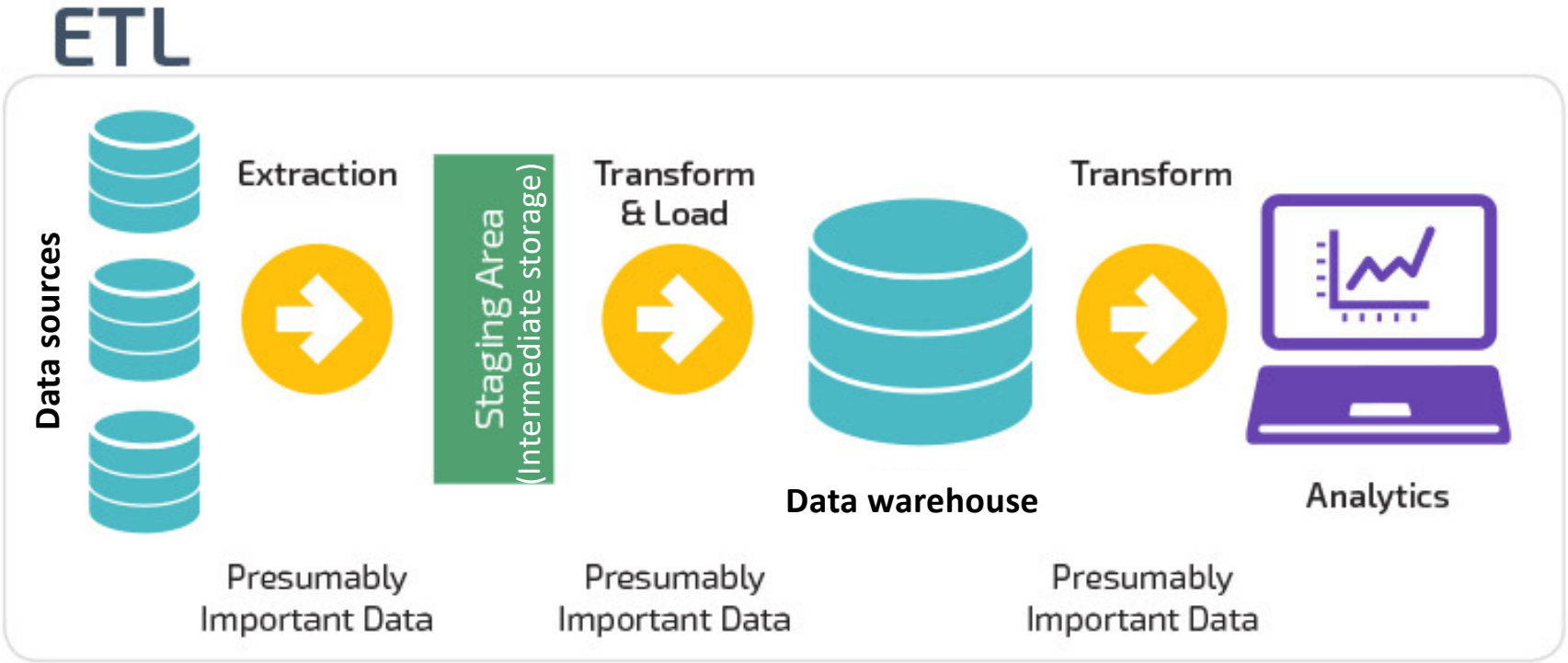
HDF5®



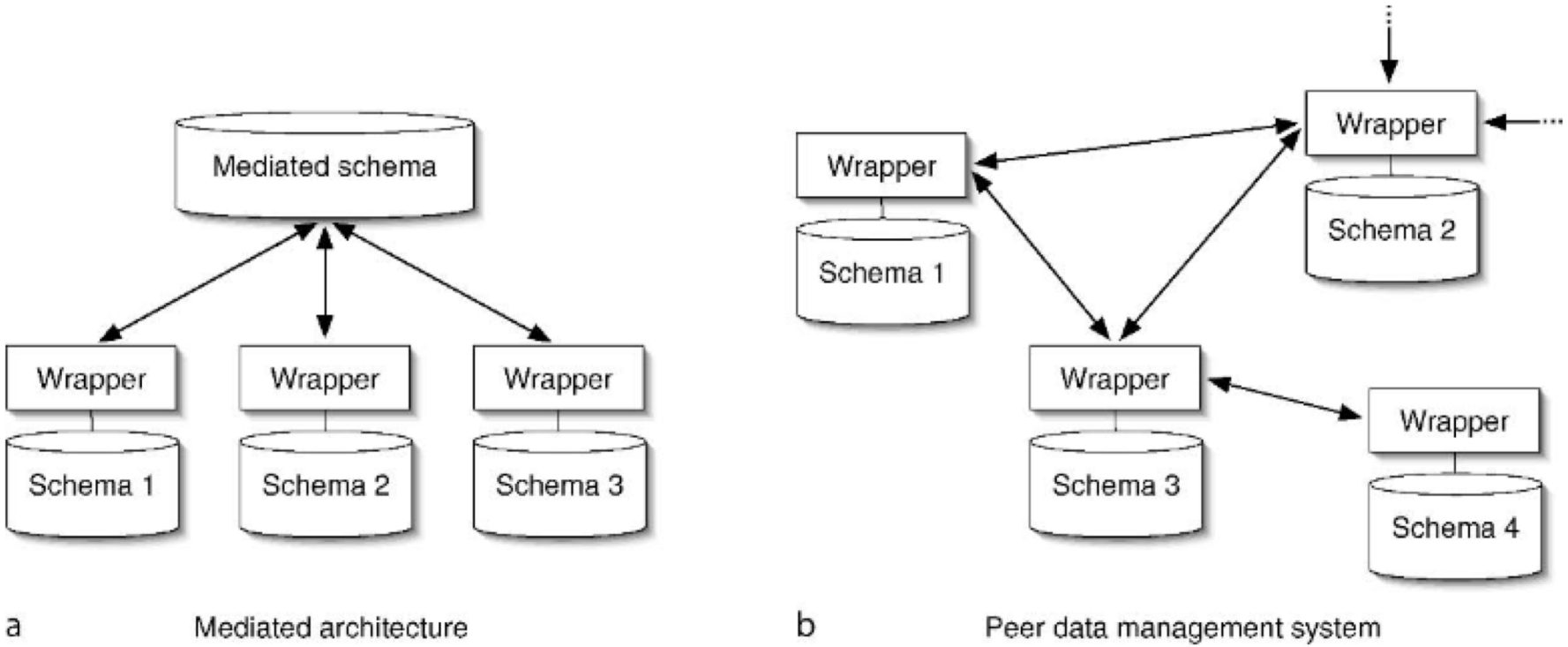
Approaches to achieve DII

Data Integration – Architecture Part 1

- DI as part of a Data warehouse architecture (data centralization)
  - **E**xtract-**T**ransform-**L**oad (ETL) operations => metadata, raw and summary data



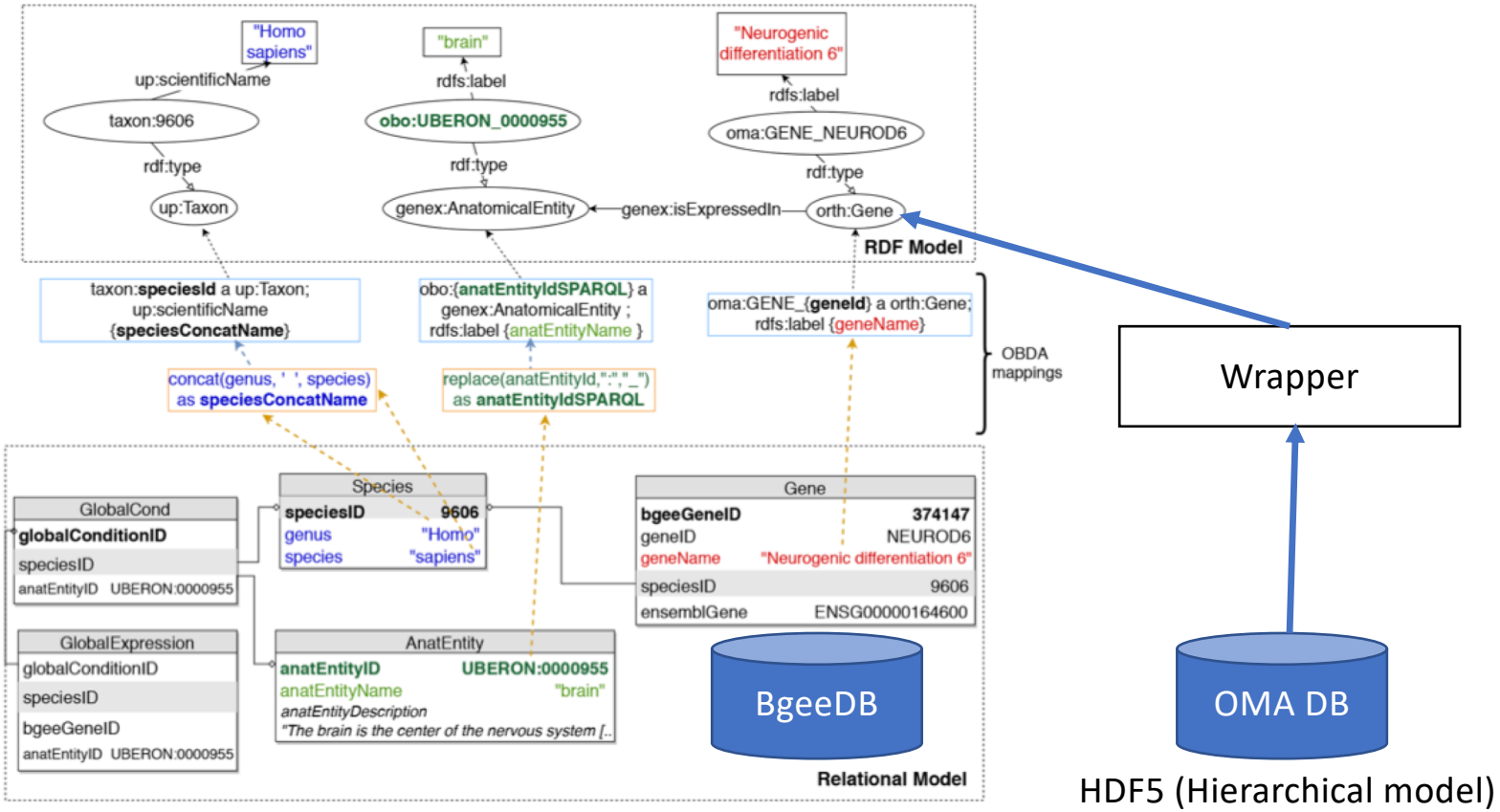
Data Integration – Architecture Part 2



~similar to Federated Database Architecture

Cudré-Mauroux P. (2009) Peer Data Management System. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA

Data Integration – Architecture Part 2



Data Integration - Mappings

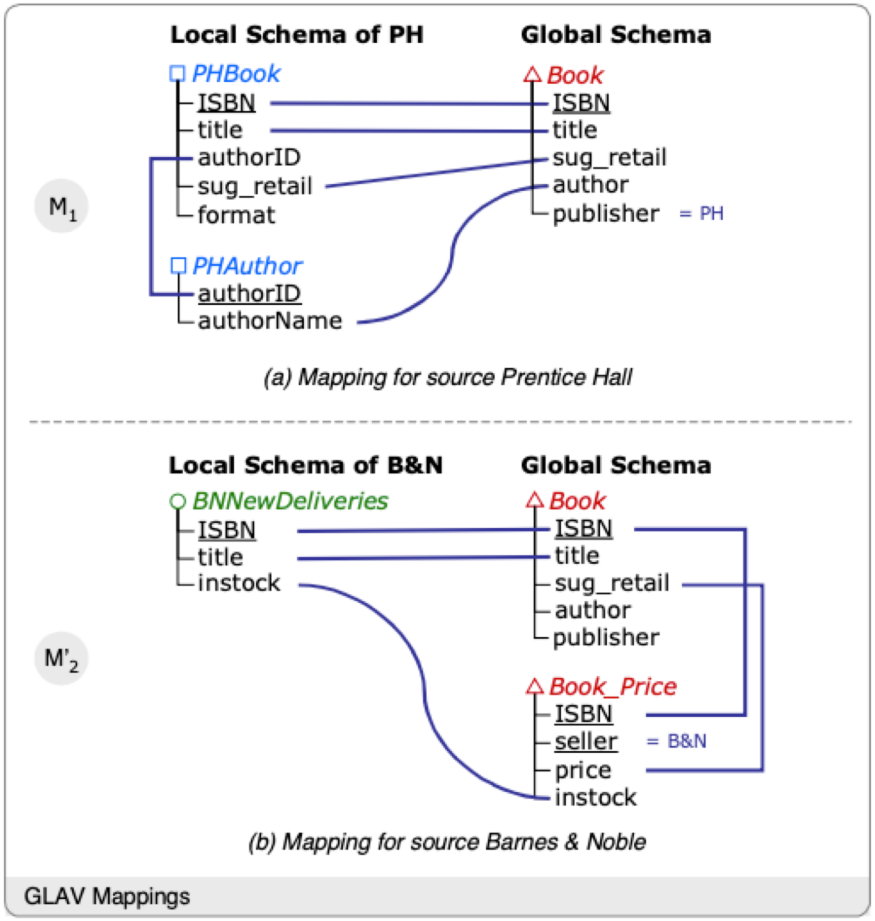
1. Global-as-View (GAV)
   - The mediated schema (i.e. global schema) is defined based on terms of the local schemas

2. Local-as-View (LAV)
   - The local schemas are described based on terms of the global schema

3. Global-Local-as-View (GLAV)
   - A superset of GAV and LAV mappings

Katsis, Y., & Papakonstantinou, Y. (2009). View-based data integration. *Encyclopedia of Database Systems*, 3332-3339.

GLAV mappings



PHBook(ISBN, title, authorID, sug_retail, format),
PHAuthor(authorID, authorName) → I(**Book**)

I(**BNNewDeliveries**) →
Book(ISBN, title, sug_retail, author, publisher),
Book_Price(ISBN, "B&N", sug_retail, instock)

Where I is the identity query : it includes all attributes of a relation (e.g. a table)

Linked data approach (on the Web Context)

1. Use **URIs** to identify things

URI_EX = http://purl.uniprot.org/taxonomy/9601

2. Use **HTTP** URIs so that people can look up those names

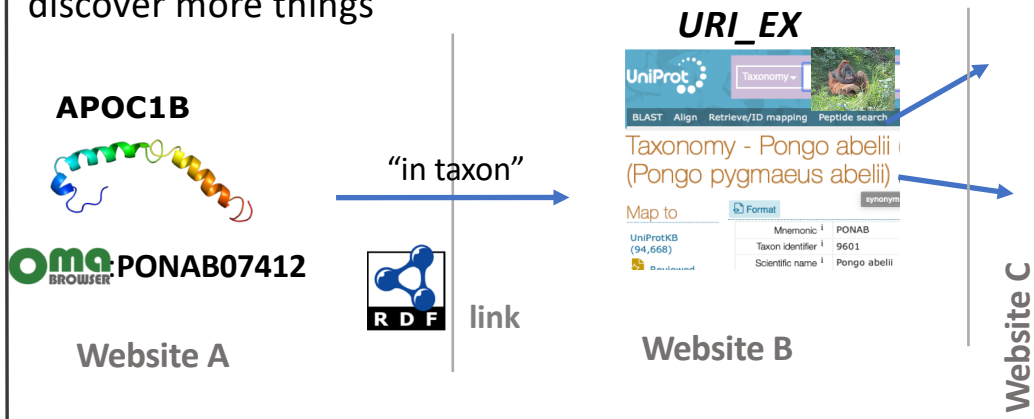
3. When someone **looks up** a URI, provide useful information, using the standards (**RDF**, RDFS, SPARQL, etc).

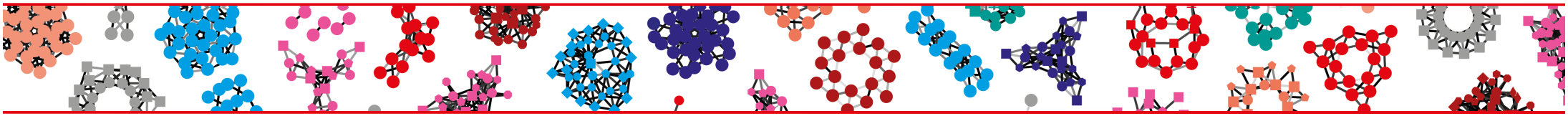
RDF Model: (Subject, Predicate, Object)

http://purl.uniprot.org/core/**commonName**

https://www.w3.org/standards/semanticweb/data.html

4. Include **links to other URIs**, so that they can discover more things





Review Plan

Data Integrations in life science: three key challenges and how to overcome them

Review plan

- **Introduction**:
  - Importance and past success of data integration
  - But still so hard to achieve DII
  - Specific approaches, wheel is reinvented!
  - Here examine why

Challenges

1. **Semantic reconciliation**
   - Tension between general vs specific (different granularities)
   - “Relaxation”
   - Examples and solutions
     - VoiDext

OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"

OTM 2019: On the Move to Meaningful Internet Systems: OTM 2019 Conferences pp 607-625 | Cite as

VoIDext: Vocabulary and Patterns for Enhancing Interoperable Datasets with Virtual Links

Authors | Authors and affiliations

Tarcisio Mendes de Farias ✉, Kurt Stockinger, Christophe Dessimoz

Challenges

2. No generic solution ("No free lunch theorem")

- Centralised vs decentralised
- How knowledge domain differences affect data integration?
- Application-specific trades-off
  - A DII approach is chosen depending on the integration problem
- Examples, solutions
  - A loosely coupled federated architecture (there is not an explicit mediated schema)

Volume 2019
2019

Enabling semantic queries across federated bioinformatics databases

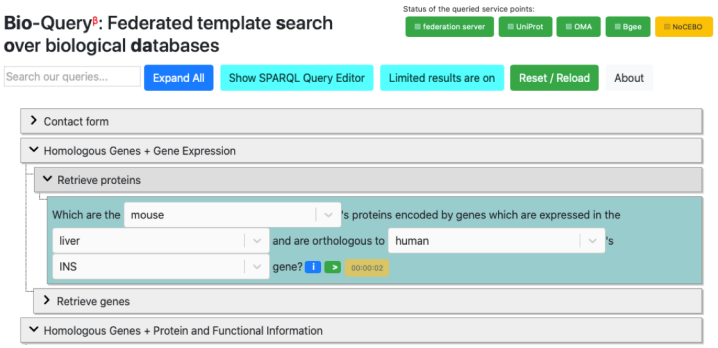
Ana Claudia Sima, Tarcisio Mendes de Farias ✉, Erich Zbinden, Maria Anisimova, Manuel Gil, Heinz Stockinger, Kurt Stockinger, Marc Robinson-Rechavi ✉, Christophe Dessimoz ✉
Author Notes

Database, Volume 2019, 2019, baz106, https://doi.org/10.1093/database/baz106
Published: 07 November 2019 **Article history** ▾

Challenges

3. Usability

- Data integration driven by usability (view as an end-to-end process)
- Flexibility and “Pareto law” (80/20)
  - Data modelling consequences
  - Technologies choice, languages, data formats, etc. (e.g.: XML vs JSON)
- Example:



SwissOrthology

swissorthology.ch

Challenges

4. **Current state** of data integration and interoperability in Life sciences
   - For structured data, data models are often missing or not following standards and might have several pitfalls and flaws
   - Good practices in terms of data modeling and ontology engineering is not always being considered

- **Discussions**
  - Is it a utopia to consider a fully automatic data integration system of general purpose? How far are we of this solution?

Recommended papers

ELSEVIER

Journal of Biomedical Informatics

Volume 41, Issue 5, October 2008, Pages 687-693

State of the nation in data integration for bioinformatics

Carole Goble, Robert Stevens

In Bioinformatics, “the *integration* of resources—a prerequisite for most bioinformatics analysis—is a perennial and costly challenge.”

Goble, C. & Stevens, R., 2008. State of the nation in data integration for bioinformatics. Journal of biomedical informatics, 41(5), pp.687–693.

Recommended papers

ELSEVIER

Journal of Biomedical Informatics

Volume 41, Issue 5, October 2008, Pages 706-716

Bio2RDF: Towards a mashup to build bioinformatics knowledge systems

François Belleau [a], Marc-Alexandre Nolin [a, b], Nicole Tourigny [b], Philippe Rigault [a], Jean Morissette [a, c]

After 6 years (last release 2014)

Bio2RDF Release 3: A Larger Connected Network of Linked Data for the Life Sciences

Michel Dumontier[1], Alison Callahan[1], Jose Cruz-Toledo[2], Peter Ansell[3], Vincent Emonet[4], François Belleau[4], Arnaud Droit[4]

[1]Stanford Center for Biomedical Informatics Research, Stanford University, CA; [2]IO Informatics, Berkeley, CA; [3]Microsoft QUT eResearch Centre, Queensland University of Technology, Australia; [4]Department of Molecular Medicine, CHUQ Research Center, Laval University, QC

Enhancing the maintainability of the Bio2RDF project using declarative mappings (2019)

Ana Iglesias-Molina, David Chaves-Fraga, Freddy Priyatna and Oscar Corcho

Recommended papers

Comment | Open Access | Published: 15 March 2016

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, [...] Barend Mons

Scientific Data **3**, Article number: 160018 (2016) | Cite this article

"Interoperability—the ability of data or tools from non-cooperating resources to integrate or work together with minimal effort."

Findable, Accessible, **Interoperable** and Reusable (FAIR)

Recommended papers

Introduction | Open Access | Published: 13 March 2014

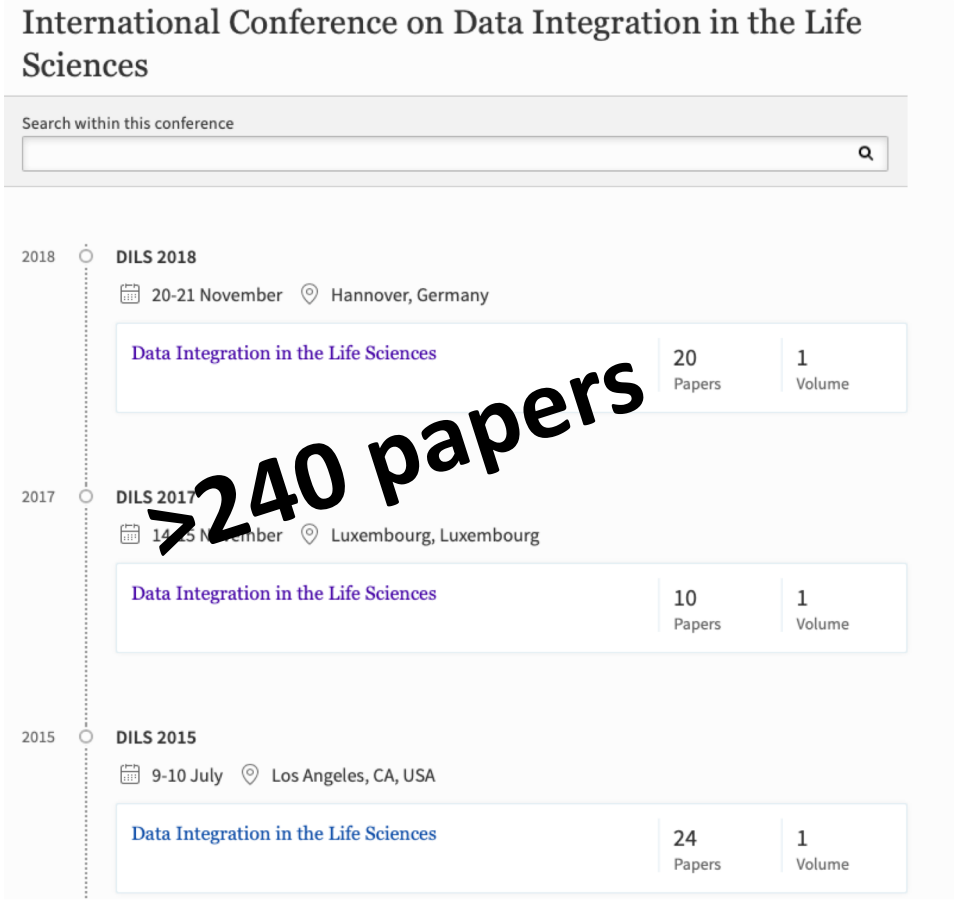
Data integration in the era of omics: current and future challenges

David Gomez-Cabrero, Imad Abugessaisa, Dieter Maier, Andrew Teschendorff, Matthias Merkenschlager, Andreas Gisel, Esteban Ballestar, Erik Bongcam-Rudloff, Ana Conesa & Jesper Tegnér

BMC Systems Biology 8, Article number: I1 (2014) | Cite this article

26k Accesses | 155 Citations | 15 Altmetric | Metrics

… the classification of data as "similar" or "heterogeneous" are still sometimes an open question which clearly depends on the specific context. Hamid and collaborators define data as similar if they are from the "same underlying source" (e.g. all gene expression) and as heterogeneous if at least two fundamentally different data sources are involved (e.g. SNP and gene expression).



>240 papers

https://link.springer.com/conference/dils

Not recommended

Lapatas, V. et al., 2015. Data integration in biological research: an overview. *Journal of biological research* , 22(1), p.9.