An overview of data integration for bioinformatics

Papers by: Carole Goble and Robert Stevens
Vaida Jakoniene and Patrick Lambrix

Presented by: David Lebech

What is data integration?

... combining data residing in different sources and providing users with a unified view of these data.

– Wikipedia[1]

… a way to retrieve data from different sources and assemble it in a unified way

– How Stuff Works[2]

Bioinformatics data sources

- 96 in 2001 versus 800+ in 2007
- Why are there so many?
  - It is easy to publish on the web
  - Hosting a resource gives reputation (?)
  - Many types of data
    - Subdisciplines develop biases
  - Bioinformatics is diverse
    - Results in specialist databases instead of central ones
  - Easier to create than reuse databases

Data source problems

- Many bioinformatics data sources are:
  - Replicating data
  - Overlapping concepts
  - Presenting different views of same type of data
- Example
  - 318 pathway databases
    - When I looked at pathguide.org while writing this

Data source problems

- Different formats
  - Flatfile, XML, Tables
- Changing interfaces, schemas and formats
- Volatile research field = volatile data sources
  - Many data sources disappear
    - As we shall see later

Data integration challenges

- Bioinformatics data has high complexity
  - Diverse sample sources
  - Variable data quality
  - Diverse types of data
  - Many interlinked collections
  - Changeable data
- How can we deal with this complexity?
  - Many more challenges need to be overcome

Data integration challenges

- Common, shared identities and names
  - E.g. WS-1 protein has ten different names
  - Determining equivalence between entities with different names is challenging
  - Need for *de facto* naming authorities
    - Heavily debated
- Shared semantics
  - Community standards for data schema and data values

Data integration challenges

- Shared, stable access mechanisms
  - Protocols, messages, interfaces, APIs
    - Stability is an issue
- Standardization
  - “Standards are boring”
- Maintaining rapid innovation and creativity

- What are the current solutions?

Service oriented architectures

- CORBA (not widely used anymore)
- Web services
    - XML/SOAP based
- Not data integration in itself
    - More focus on the interface than the data integration

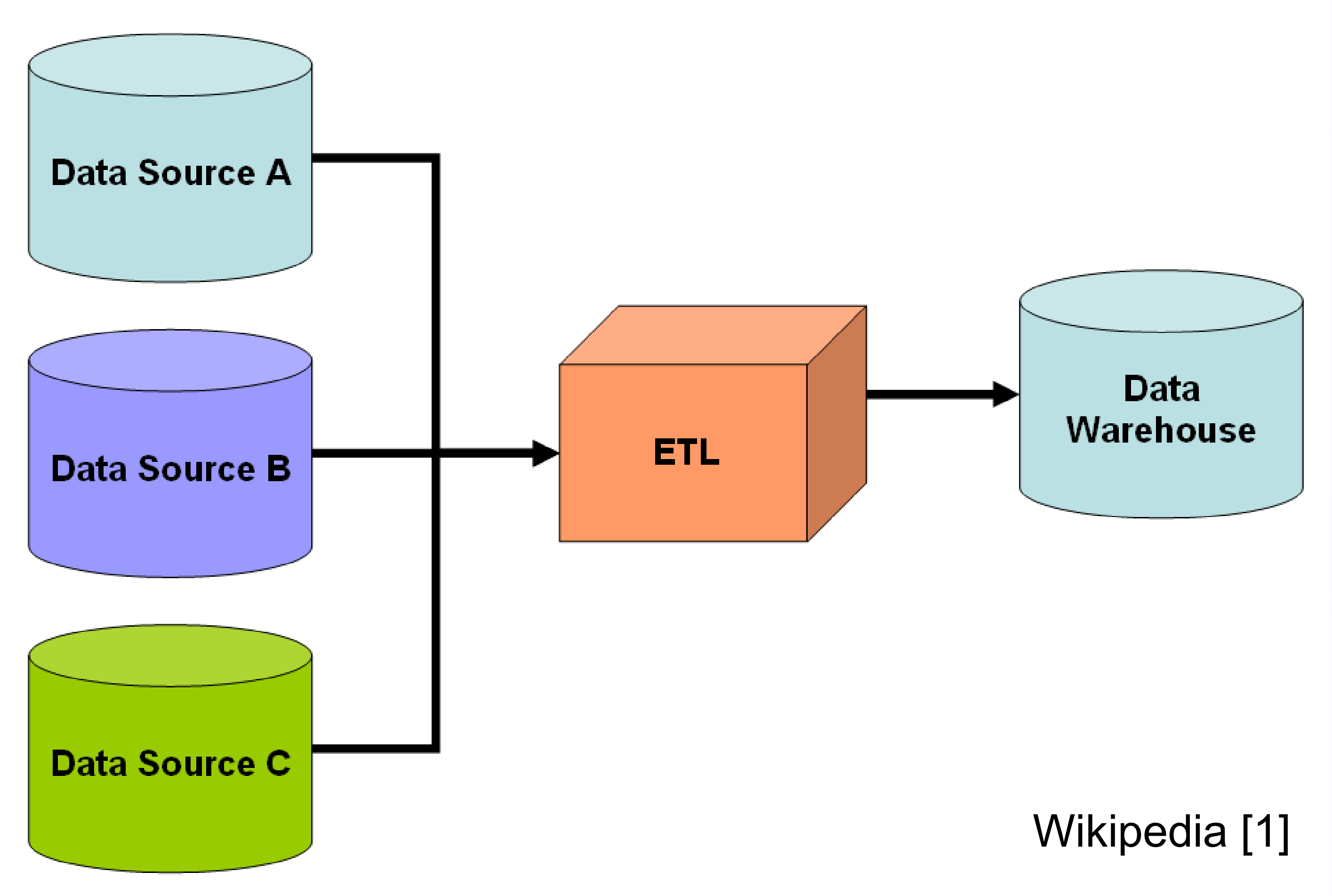
Link integration

- Cross-referencing data entries in different data sources
- Often relies on ontologies
- Most effective and widely used
- Problems:
  - Vulnerable to name crashes
  - Model dependent
  - Not real integration

Data warehousing

- Single, integrated resource
- Requires a pre-determined model
  - Extract, cleans and integrates multiple data sources into the model
- Problems:
  - Expensive
  - Fixed or hard to change
  - Not flexible
    - “Data mortuaries”

Data warehousing

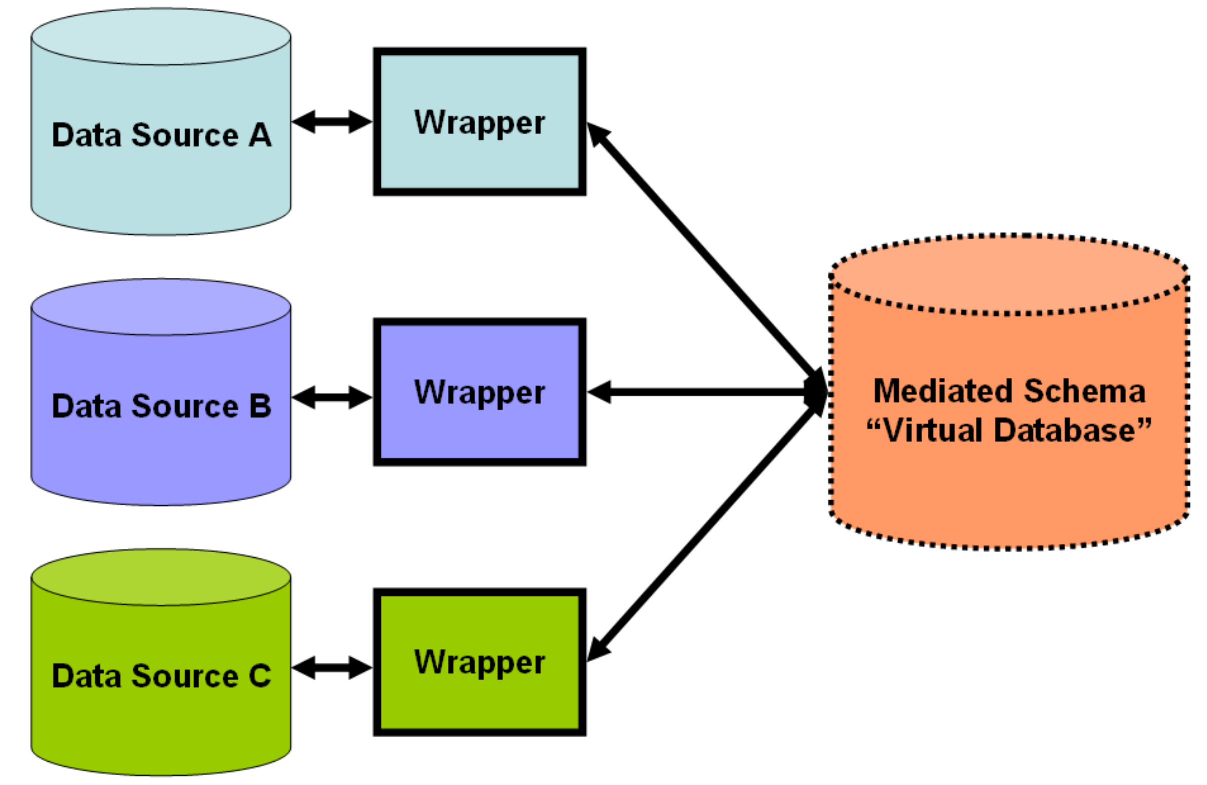


Wikipedia [1]

View integration

- Data is kept at original source
- A “virtual warehouse” is set up so it looks like one data source
- Contents are always new (or “fresh”)
- Problems:
  - Expensive to set up model
  - Complex

View integration

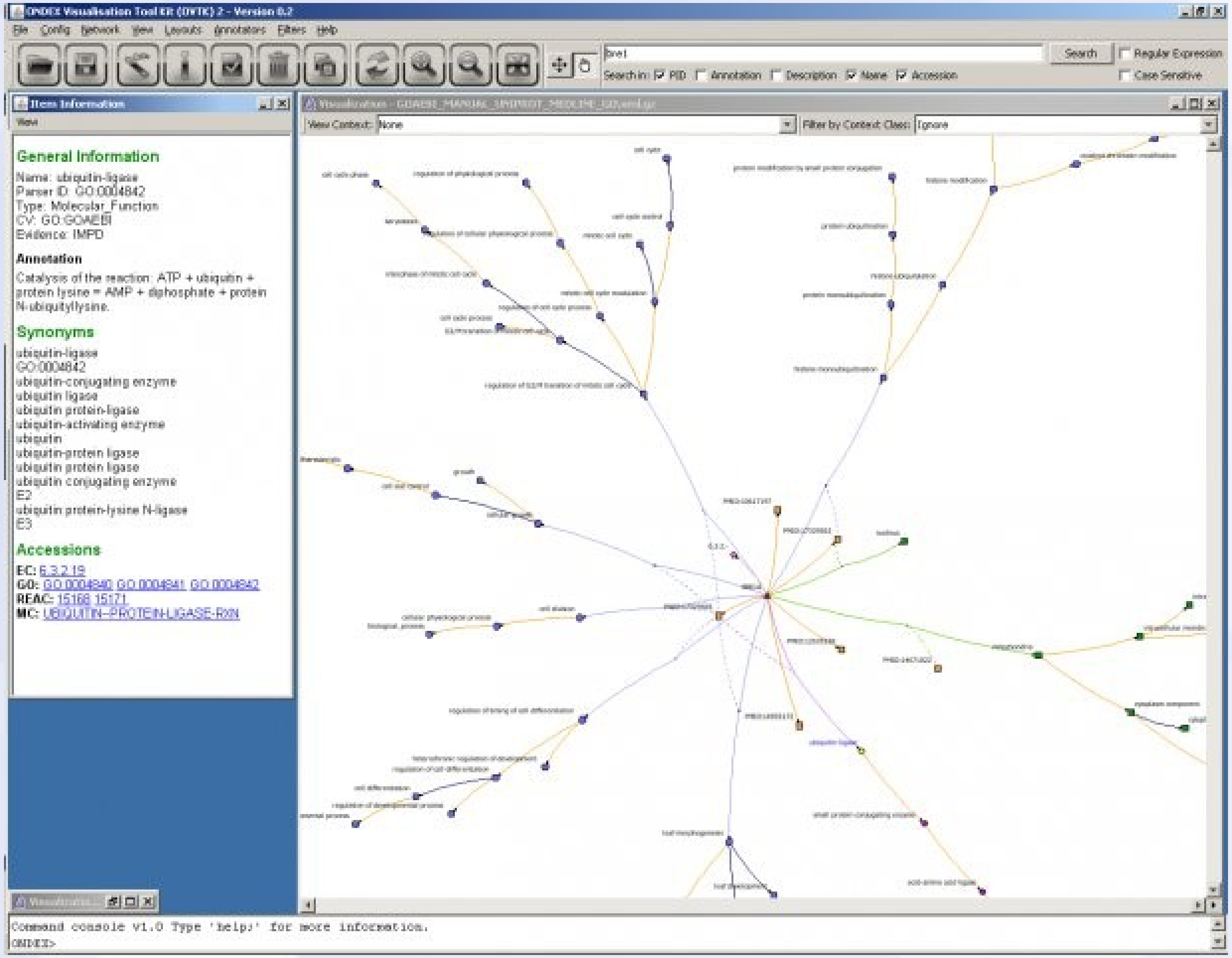


Wikipedia [1]

Integration applications

- Specific integration systems designed for a single domain
  - E.g. Ensembl, Toolbus, Utopia, Ondex
- Single application domain
- Some apps use “Workflows”
  - “the coordination of one or more services into a data analysis pipeline”
    - Taverna FAQ [3]

Integration application – Ondex


ONDEX Visualisation Tool Kit (OVTK) 2 - Version 0.2
File Config Network View Layouts Annotators Filters Help
brel
Search
Search in: ☑ PID □ Annotation □ Description ☑ Name ☑ Accession
□ Regular Expression
□ Case Sensitive
Item Information
View
Visualisation - GOAEBI_MANUAL_UNIPROT_MEDLINE_GO.xml.gz
View Context: None
Filter by Context Class: Ignore
Visualisatio...

General Information

Name: ubiquitin-ligase
Parser ID: GO:0004842
Type: Molecular_Function
CV: GO:GOAEBI
Evidence: IMPD

Annotation

Catalysis of the reaction: ATP + ubiquitin + protein lysine = AMP + diphosphate + protein N-ubiquityllysine.

Synonyms

ubiquitin-ligase
GO:0004842
ubiquitin-conjugating enzyme
ubiquitin ligase
ubiquitin protein-ligase
ubiquitin-activating enzyme
ubiquitin
ubiquitin-protein ligase
ubiquitin protein ligase
ubiquitin conjugating enzyme
E2
ubiquitin protein-lysine N-ligase
E3

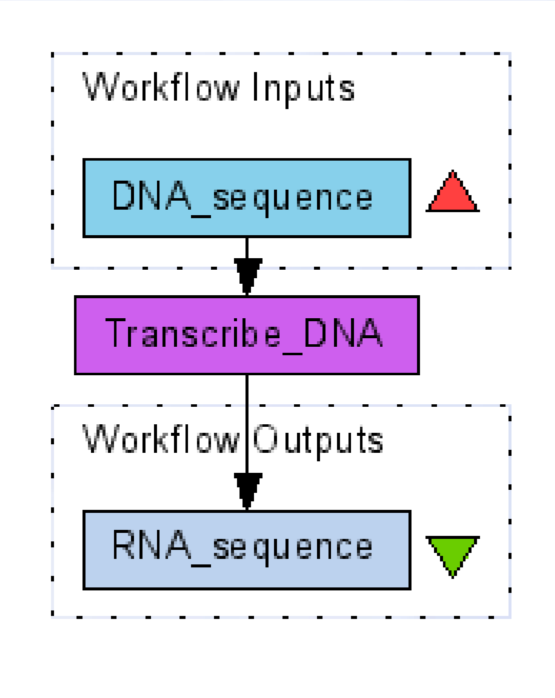
Accessions

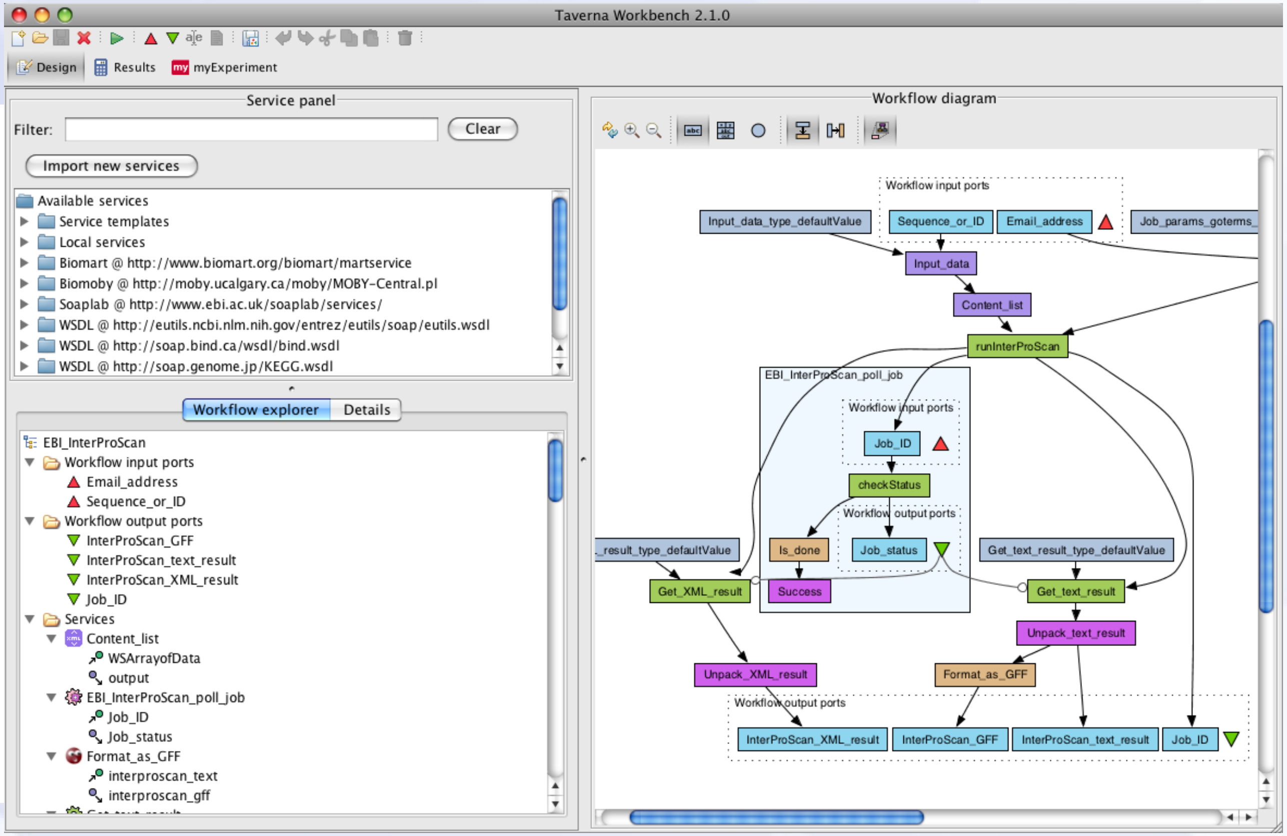
EC: 6.3.2.19
GO: GO:0004840 GO:0004841 GO:0004842
REAC: 15168 15171
MC: UBIQUITIN--PROTEIN-LIGASE-RXN

```
Command console v1.0 Type 'help;' for more information.
ONDEX>
```

Protein annotated to the Gene Ontology [5]

Integration application – Taverna





Taverna [3]

Mashups

- Combining data from different resources to give new views in the data
  - E.g. Tracking the flu through Google Earth
- Open, light and on-demand
  - “just-in-time, just enough”
- Aggregation rather than integration
- Transient
- Problems:
  - Same as for link integration

The future

- Web 2.0 and The Semantic Web are promising for the bioinformatics field
  - Datasets as RDF documents
  - OWL as ontology language
  - Goal is to create a single Web of biological data
- “Conclusion”:
  - Focus on the identity and naming problem

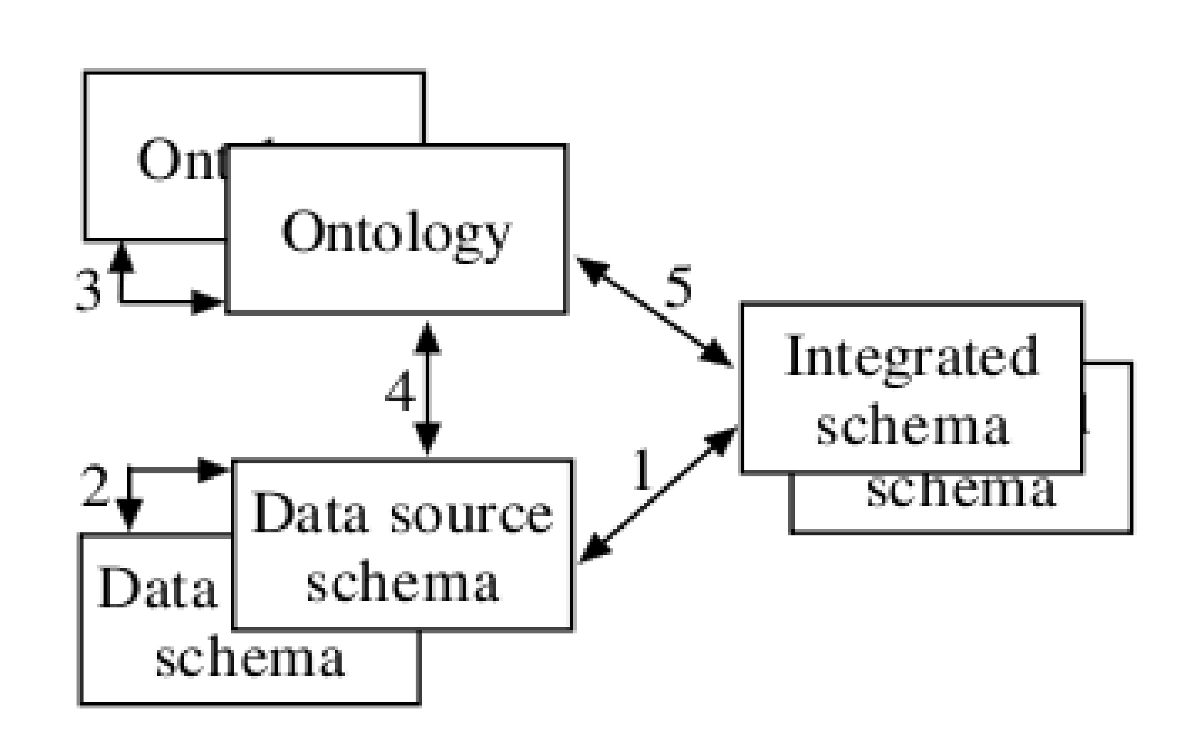
Ontology-based data integration

- Paper by Jakoniene and Lambrix
- Presents many of the same issues as I have just outlined
- Proposes an information integration system (IIS) for using ontology based bioinformatics data
  - I will briefly discuss this part

Proposal

- Integration of data sources
    - Global-as-view, local-as-view
- Cross-referencing between data sources
- Ontological alignment
- Ontology-based data source descriptions
- Ontology-based integrated schemas

A pretty chart



“Current” approaches

- BACIIS, KIND, SEMEDA
    - No longer available
        - None of them
- TAMBIS
    - No longer available
        - “We could ask good questions, but the underlying resources couldn't answer them.”
            - Robert Stevens on an ontology forum [4]
                - He is one of the researchers behind Tambis

What's wrong with this picture?

- The paper is written in 2005
  - Perfect example of current issues
    - Bioinformatics is a rapidly changing field
    - Data sources are rapidly changing
- “Future Work”
  - BioTRIFU
    - Last publication came out in 2005
    - Nothing available at the website
- In my opinion, this paper is a bit embarrassing

Conclusion

- Data integration is important for bioinformatics research
- Heterogeneous data is difficult to integrate
- Seemingly more challenges than solutions
- Need for unified naming conventions
- Fast, simple and straightforward approaches (Web 2.0 services, Semantic web) are currently gaining popularity

Thank you

Questions?

References

[1] http://en.wikipedia.org/wiki/Data_integration

[2] http://communication.howstuffworks.com/data-integration.htm

[3] http://www.taverna.org.uk/introduction/what-is-taverna/

[4] http://ontolog.cim3.net/forum/ontolog-forum/2007-08/msg00426.html

[5] http://sourceforge.net/project/screenshots.php?group_id=112544&ssid=89705

Papers:

Carole Goble and Robert Stevens State of the nation in data integration for bioinformatics. Journal of Biomedical Informatics 41, 687-693, 2008

Vaida Jakoniene and Patrick Lambrix Ontology-based integration for bioinformatics. In Proceedings of VLDB Workshop on Ontologies-based techniques for DataBases, 2005