Data integration perspectives from the LTB project

Michele Pasin

Centre for Computing in the Humanities

Kings College, London

michele.pasin@ kcl.ac.uk

SDH-SEMI-2010
Montreal, Canada,
June 2010

Summary

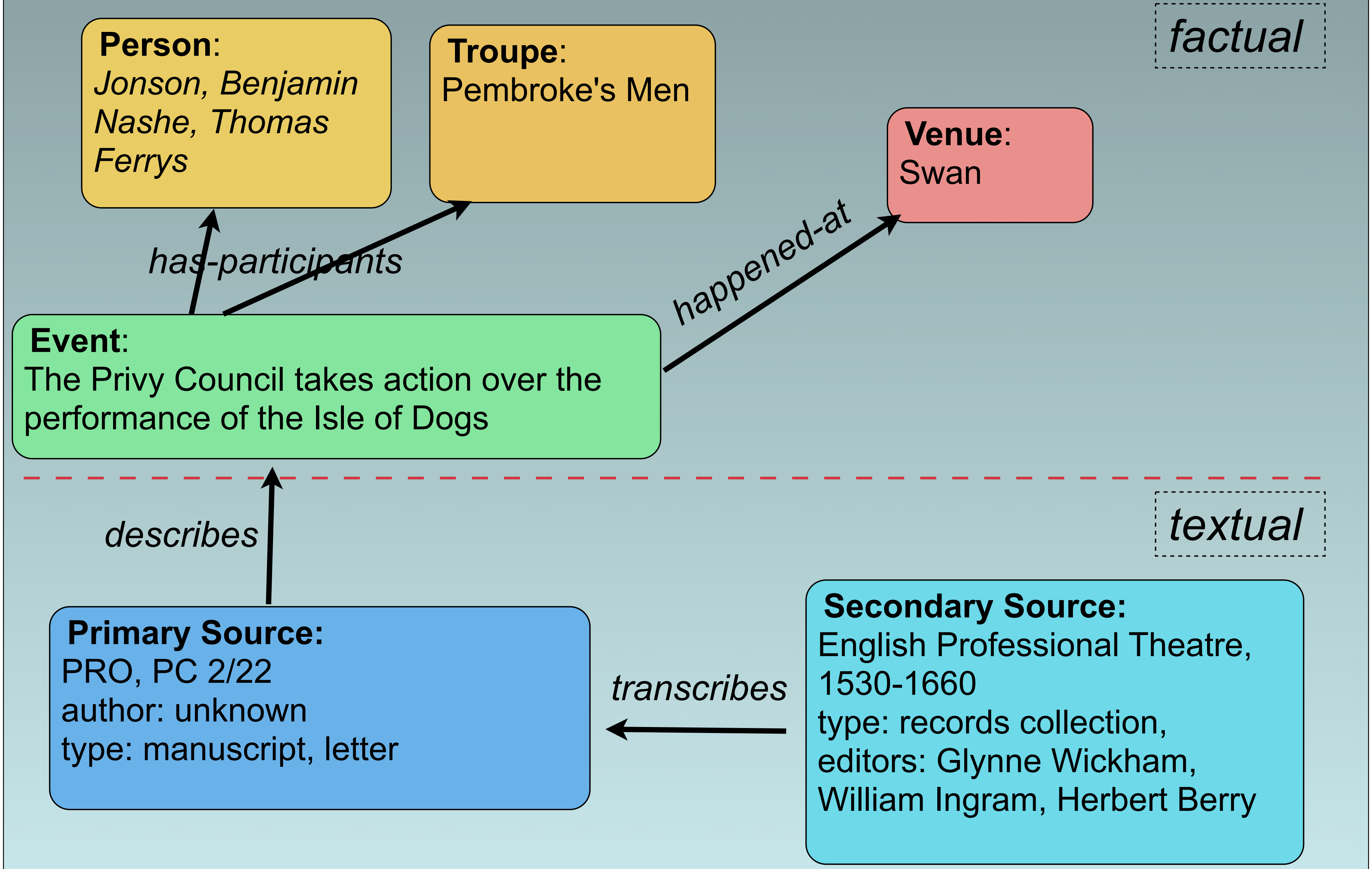
1. Why data integration? Some possible scenarios..

2. How? Existing approaches

3. Linking RDF data, a quick overview

4. Open problems, lessons learned..

The LTB data model in a nutshell



factual

Person:
Jonson, Benjamin
Nashe, Thomas
Ferrys

Troupe:
Pembroke's Men

Venue:
Swan

has-participants

happened-at

Event:
The Privy Council takes action over the performance of the Isle of Dogs

textual

describes

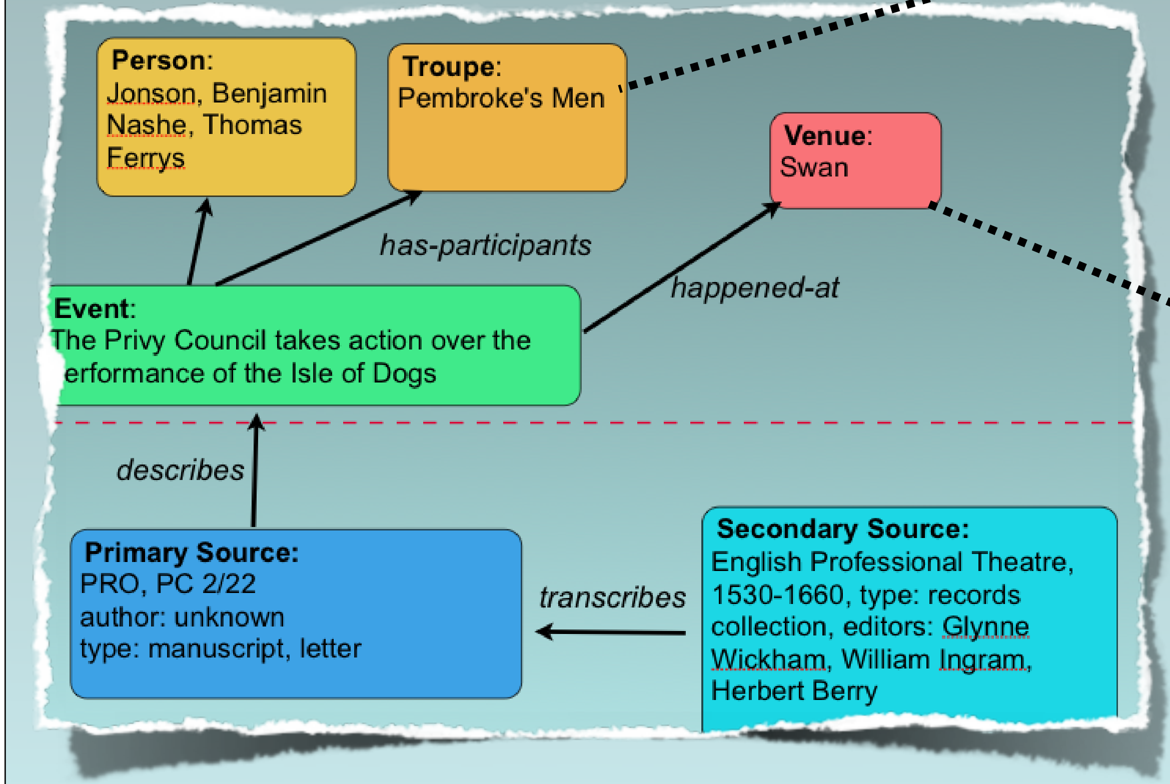
Primary Source:
PRO, PC 2/22
author: unknown
type: manuscript, letter

transcribes

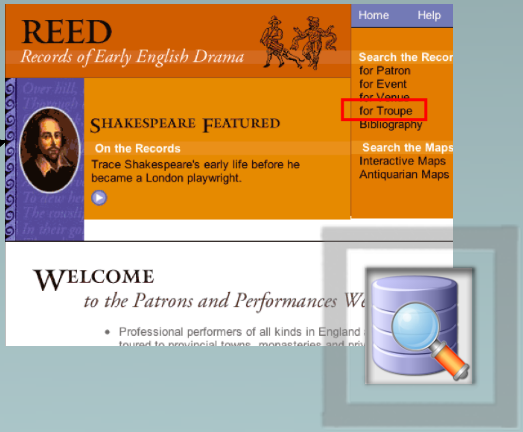
Secondary Source:
English Professional Theatre, 1530-1660
type: records collection,
editors: Glynne Wickham, William Ingram, Herbert Berry

LTB: connecting to other data (I)

- benefit from other **authoritative** sources
- integrating data which are not the focus of the project
- **building on** other projects results



P&P



REED

Records of Early English Drama

Home Help

Search the Recor
for Patron
for Event
for Venue
for Troupe
Bibliography

Search the Maps
Interactive Maps
Antiquarian Maps

SHAKESPEARE FEATURED

On the Records

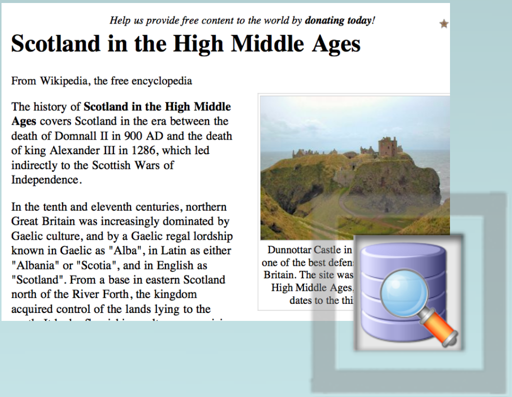
Trace Shakespeare's early life before he became a London playwright.

WELCOME

to the Patrons and Performances We

- Professional performers of all kinds in England toured to provincial towns, monasteries and priv

Geonames.org



*Help us provide free content to the world by **donating today**!*

Scotland in the High Middle Ages

From Wikipedia, the free encyclopedia

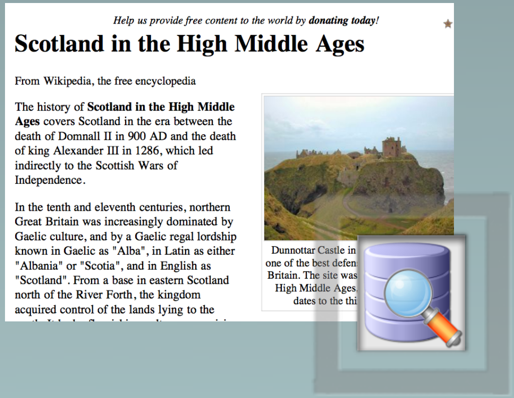
The history of **Scotland in the High Middle Ages** covers Scotland in the era between the death of Domnall II in 900 AD and the death of king Alexander III in 1286, which led indirectly to the Scottish Wars of Independence.

In the tenth and eleventh centuries, northern Great Britain was increasingly dominated by Gaelic culture, and by a Gaelic regal lordship known in Gaelic as "Alba", in Latin as either "Albania" or "Scotia", and in English as "Scotland". From a base in eastern Scotland north of the River Forth, the kingdom acquired control of the lands lying to the

Dunnottar Castle in
one of the best defen
Britain. The site was
High Middle Ages,
dates to the thi

LTB: connecting to other data (II)

DBPedia



Help us provide free content to the world by **donating today**!

Scotland in the High Middle Ages

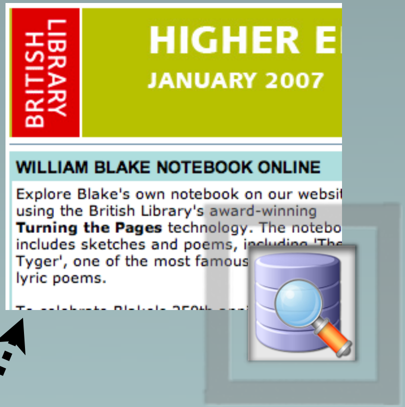
From Wikipedia, the free encyclopedia

The history of **Scotland in the High Middle Ages** covers Scotland in the era between the death of Domnall II in 900 AD and the death of king Alexander III in 1286, which led indirectly to the Scottish Wars of Independence.

In the tenth and eleventh centuries, northern Great Britain was increasingly dominated by Gaelic culture, and by a Gaelic regal lordship known in Gaelic as "Alba", in Latin as either "Albania" or "Scotia", and in English as "Scotland". From a base in eastern Scotland north of the River Forth, the kingdom acquired control of the lands lying to the

Dunnottar Castle in
one of the best defen
Britain. The site was
High Middle Ages
dates to the thi

BL collection



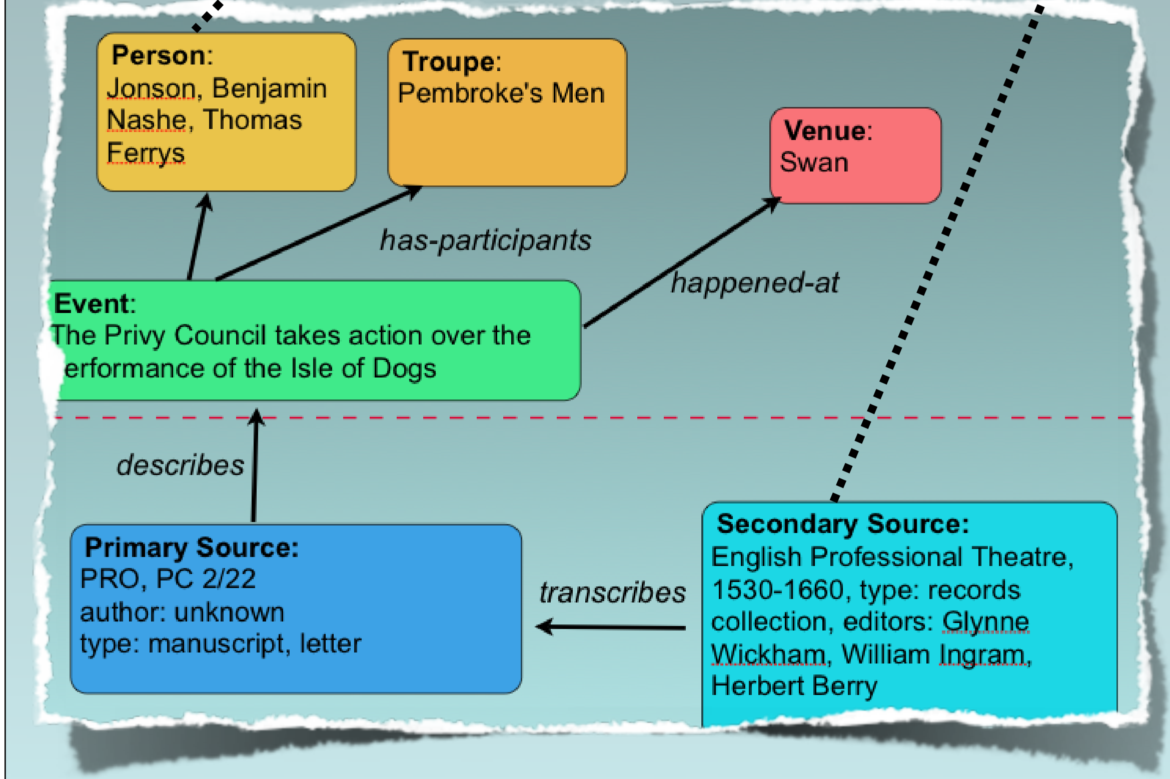
BRITISH LIBRARY

HIGHER E

JANUARY 2007

WILLIAM BLAKE NOTEBOOK ONLINE

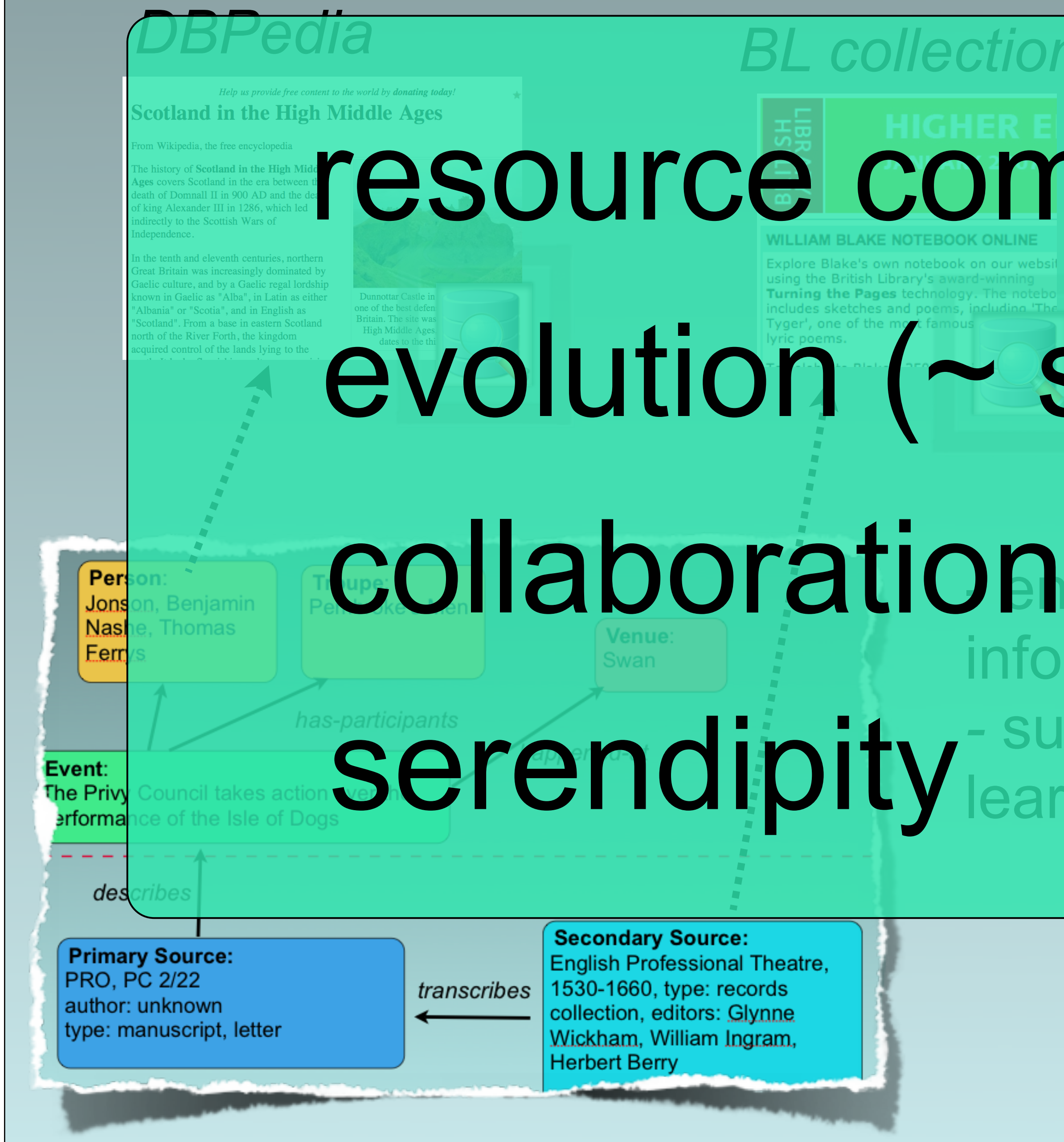
Explore Blake's own notebook on our websit using the British Library's award-winning **Turning the Pages** technology. The notebo includes sketches and poems, including 'The Tyger', one of the most famous lyric poems.



- enriching the resource with other **contextual** information, *for free*

- supporting the creation of **learning** pathways

Connecting to other data: keywords


DBPedia
BL collection
Help us provide free content to the world by donating today!
Scotland in the High Middle Ages
From Wikipedia, the free encyclopedia
HIGHER E
WILLIAM BLAKE NOTEBOOK ONLINE
Person:
Jonson, Benjamin
Nashe, Thomas
Ferrys
Troupe:
Venue:
Swan
has-participants
Event:
The Privy Council takes action over the performance of the Isle of Dogs
describes
Primary Source:
PRO, PC 2/22
author: unknown
type: manuscript, letter
transcribes
Secondary Source:
English Professional Theatre, 1530-1660, type: records collection, editors: Glynne Wickham, William Ingram, Herbert Berry

The history of Scotland in the High Middle Ages covers Scotland in the era between the death of Domnall II in 900 AD and the de[ath] of king Alexander III in 1286, which led indirectly to the Scottish Wars of Independence.

In the tenth and eleventh centuries, northern Great Britain was increasingly dominated by Gaelic culture, and by a Gaelic regal lordship known in Gaelic as "Alba", in Latin as either "Albania" or "Scotia", and in English as "Scotland". From a base in eastern Scotland north of the River Forth, the kingdom acquired control of the lands lying to the

Dunnottar Castle in ... one of the best defen... Britain. The site was ... High Middle Ages ... dates to the thi...

Explore Blake's own notebook on our websit... using the British Library's award-winning **Turning the Pages** technology. The notebo... includes sketches and poems, including 'The Tyger', one of the mo[st] famous lyric poems.

resource composition

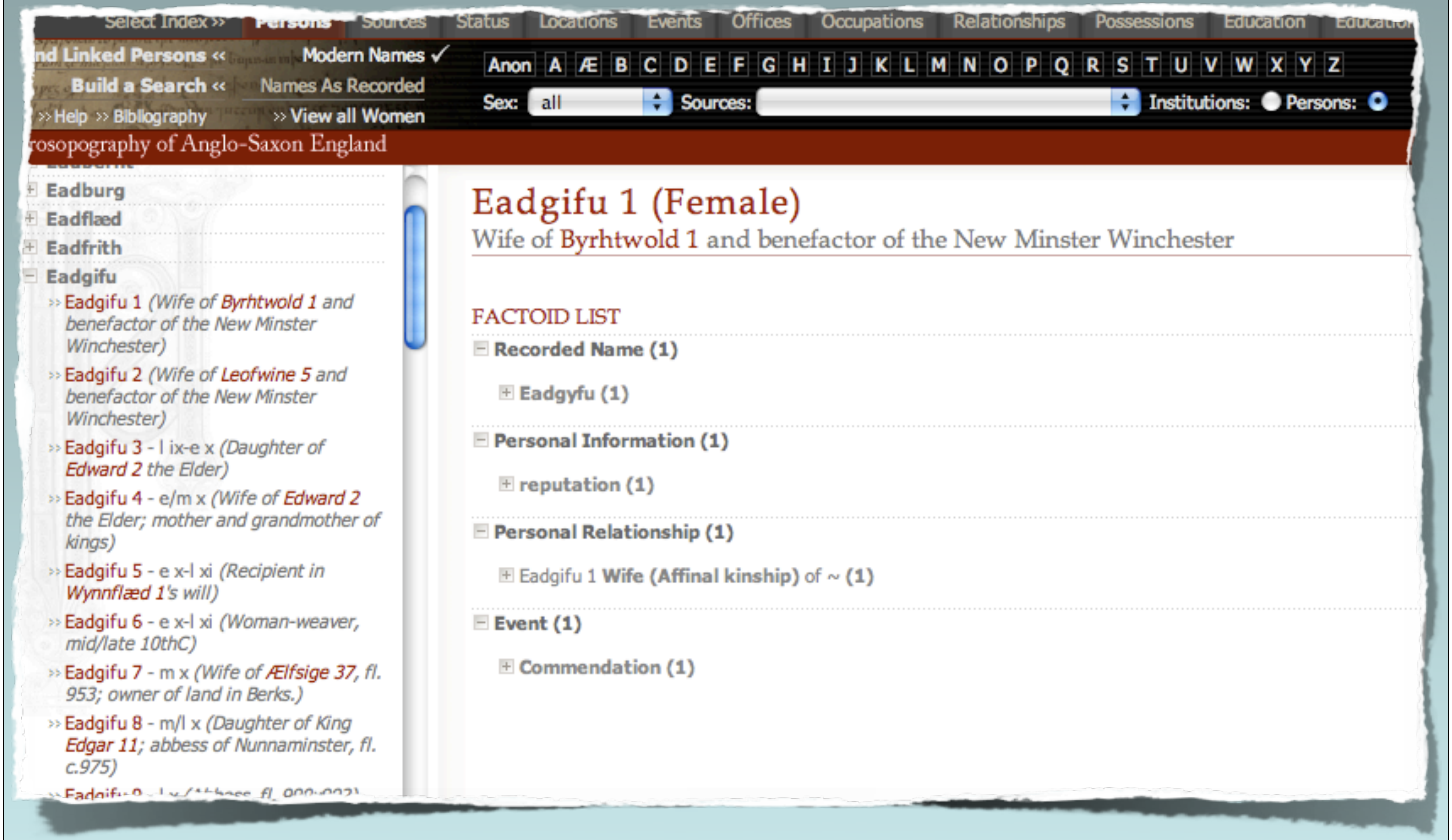
evolution (~ science)

collaboration

serendipity

- enriching the contextual information available, *for free* - supporting the creation of learning pathways

A typical DB front end..



⊞ Eadburg

⊞ Eadflæd

⊞ Eadfrith

⊟ Eadgifu

- » Eadgifu 1 *(Wife of Byrhtwold 1 and benefactor of the New Minster Winchester)*
- » Eadgifu 2 *(Wife of Leofwine 5 and benefactor of the New Minster Winchester)*
- » Eadgifu 3 - l ix-e x *(Daughter of Edward 2 the Elder)*
- » Eadgifu 4 - e/m x *(Wife of Edward 2 the Elder; mother and grandmother of kings)*
- » Eadgifu 5 - e x-l xi *(Recipient in Wynnflæd 1's will)*
- » Eadgifu 6 - e x-l xi *(Woman-weaver, mid/late 10thC)*
- » Eadgifu 7 - m x *(Wife of Ælfsige 37, fl. 953; owner of land in Berks.)*
- » Eadgifu 8 - m/l x *(Daughter of King Edgar 11; abbess of Nunnaminster, fl. c.975)*

Eadgifu 1 (Female)

Wife of Byrhtwold 1 and benefactor of the New Minster Winchester

FACTOID LIST

⊟ **Recorded Name (1)**

⊞ **Eadgyfu (1)**

⊟ **Personal Information (1)**

⊞ **reputation (1)**

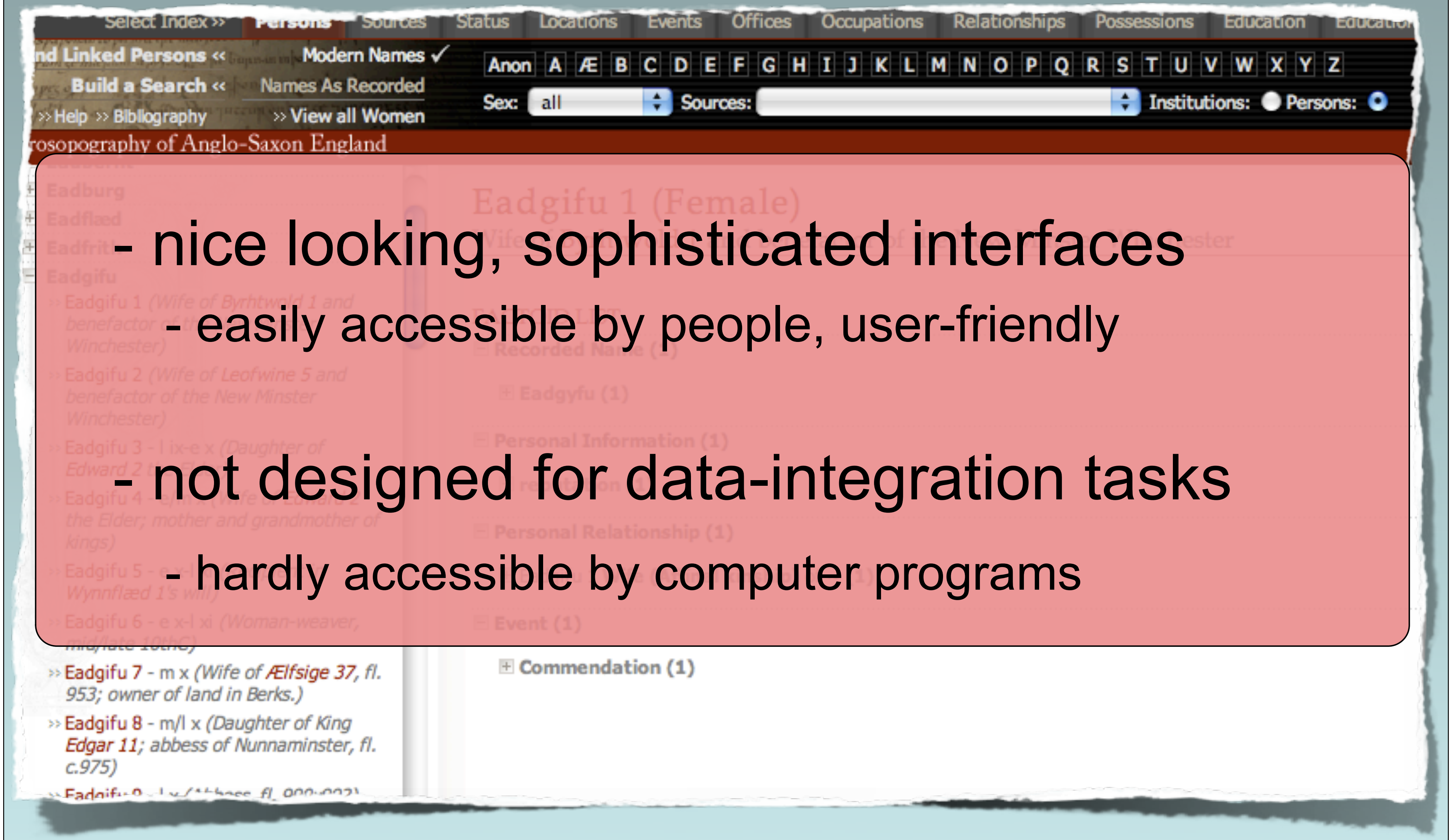
⊟ **Personal Relationship (1)**

⊞ Eadgifu 1 **Wife (Affinal kinship)** of ~ **(1)**

⊟ **Event (1)**

⊞ **Commendation (1)**

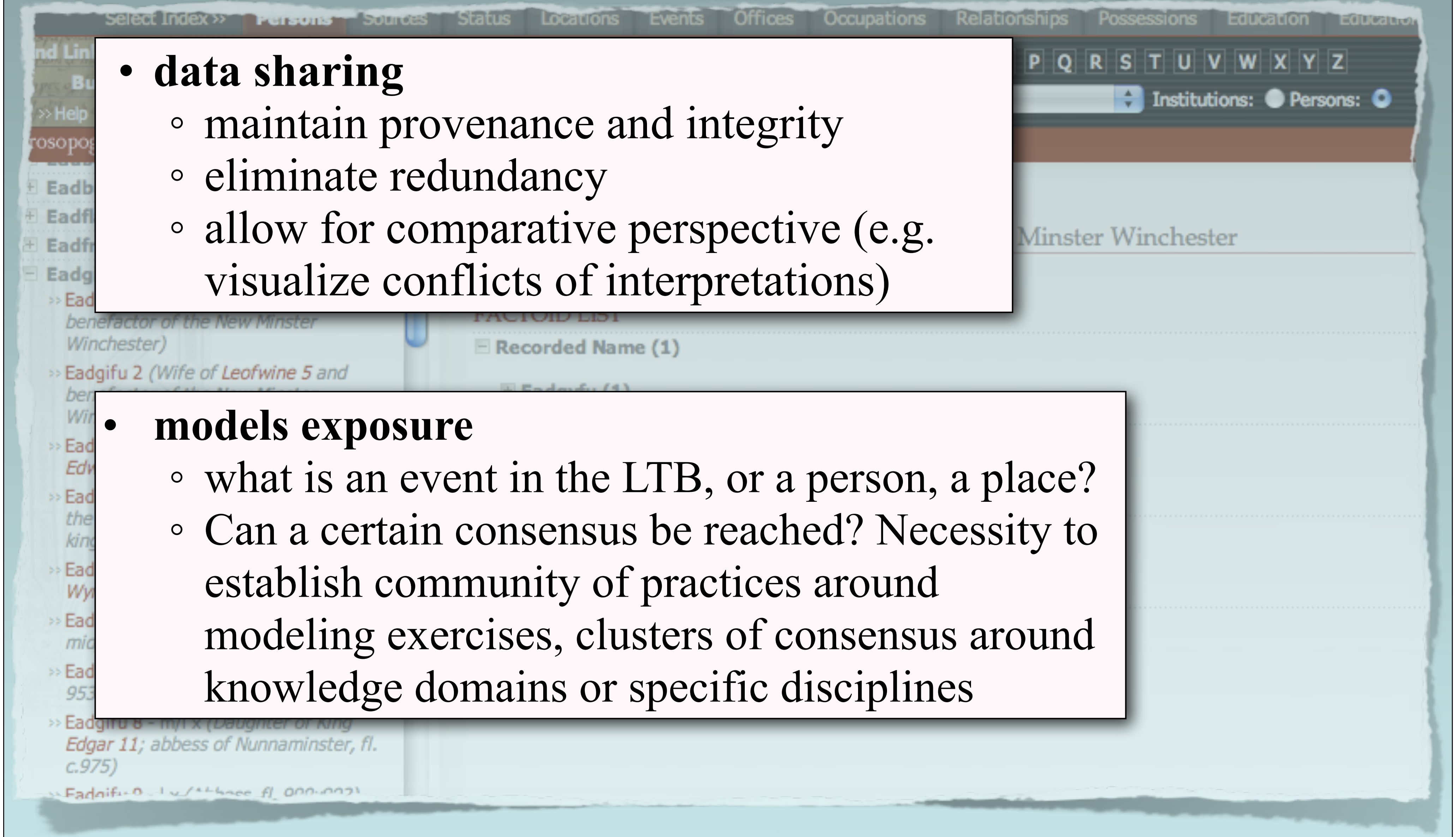
The black box effect



- nice looking, sophisticated interfaces
    - easily accessible by people, user-friendly

- not designed for data-integration tasks
    - hardly accessible by computer programs

Opening up black boxes: Desiderata



- **data sharing**
  - maintain provenance and integrity
  - eliminate redundancy
  - allow for comparative perspective (e.g. visualize conflicts of interpretations)

- **models exposure**
  - what is an event in the LTB, or a person, a place?
  - Can a certain consensus be reached? Necessity to establish community of practices around modeling exercises, clusters of consensus around knowledge domains or specific disciplines

Approaches to the integration problem

1) **database-oriented solutions**

- effective, but often not geared for open, web-oriented scenarios
- often costly and geared for enterprise use (eg IBM DB2 server)

2) **web APIs, webServices**

- eg. eBay, Amazon, Delicious, simple data feeds in XML, JSON
- query mechanisms, if present, are pretty poor / no control on feed structure

3) **Semantic Web / Linked Data solution**

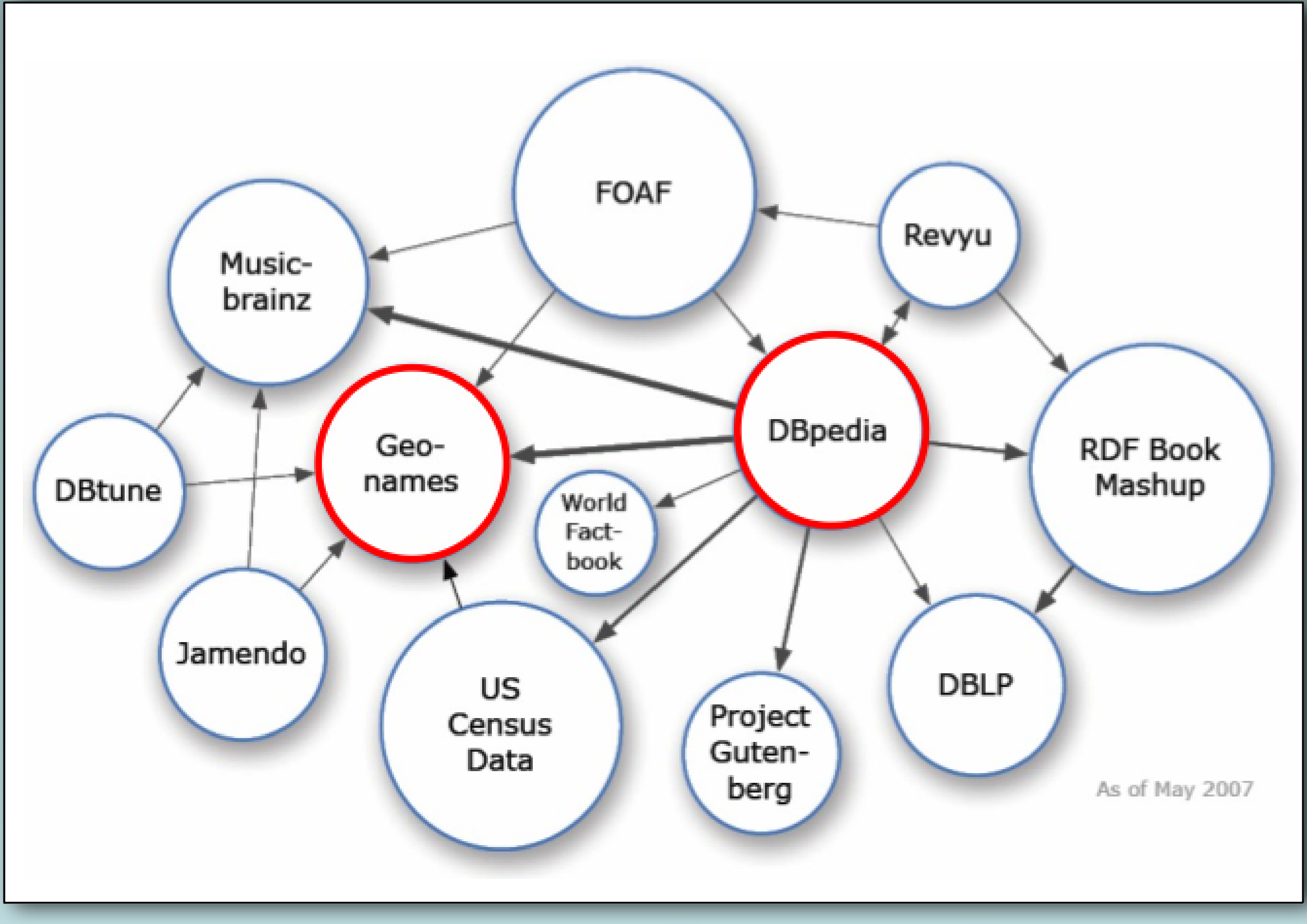
- set of standards for encoding the semantics of your data (RDF, RDFS, OWL)
- allow reasoning tasks (eg inferences)
- provide a single, standardized query mechanism (SPARQL)

4) **Hybrid approaches**

- keep working in DB as usual, with ‘on the fly’ translation to RDF
- very interesting, but still in CS research phase

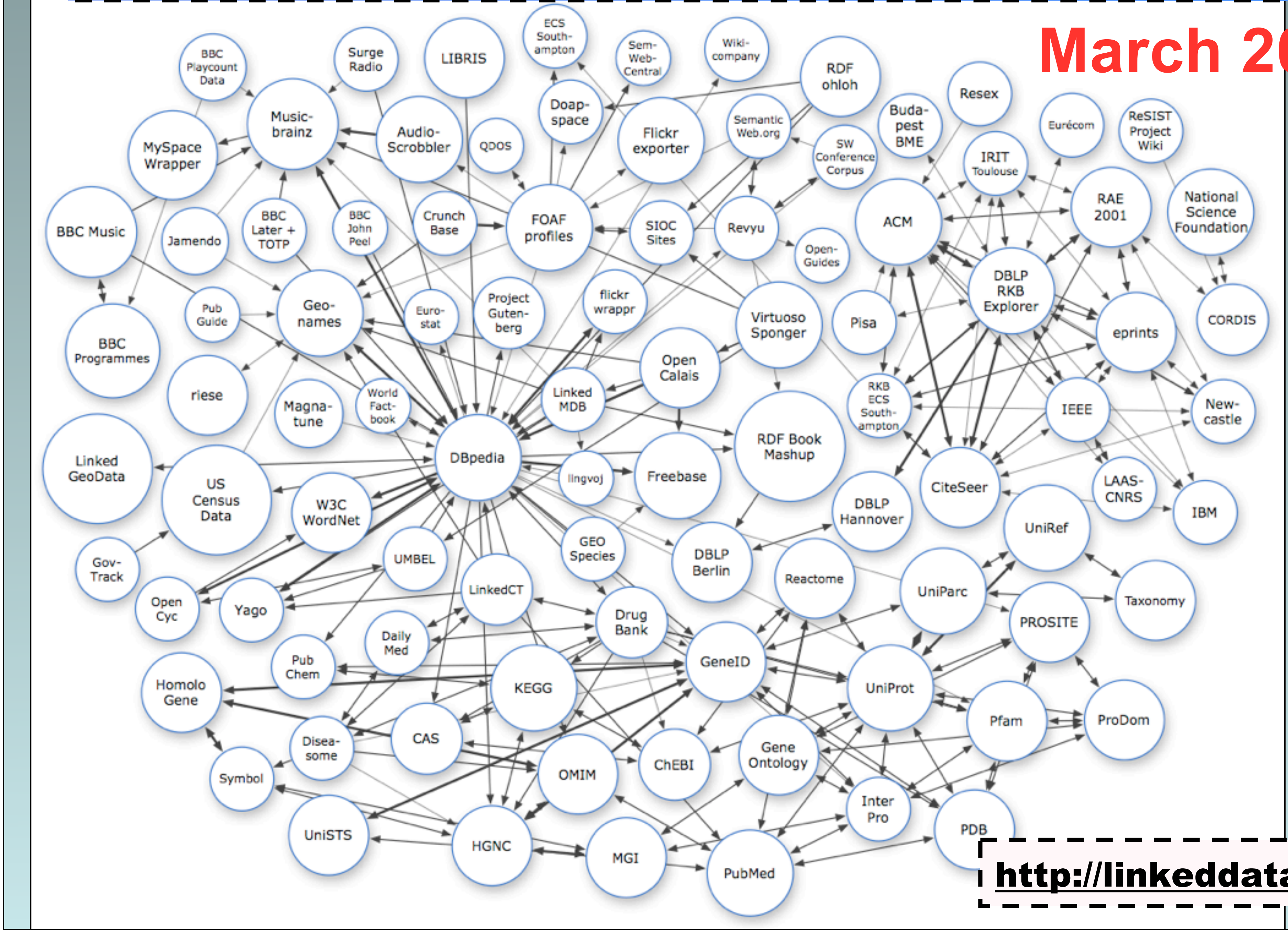
Approaches: the Linked-Data initiative

May 2007



http://linkeddata.org/

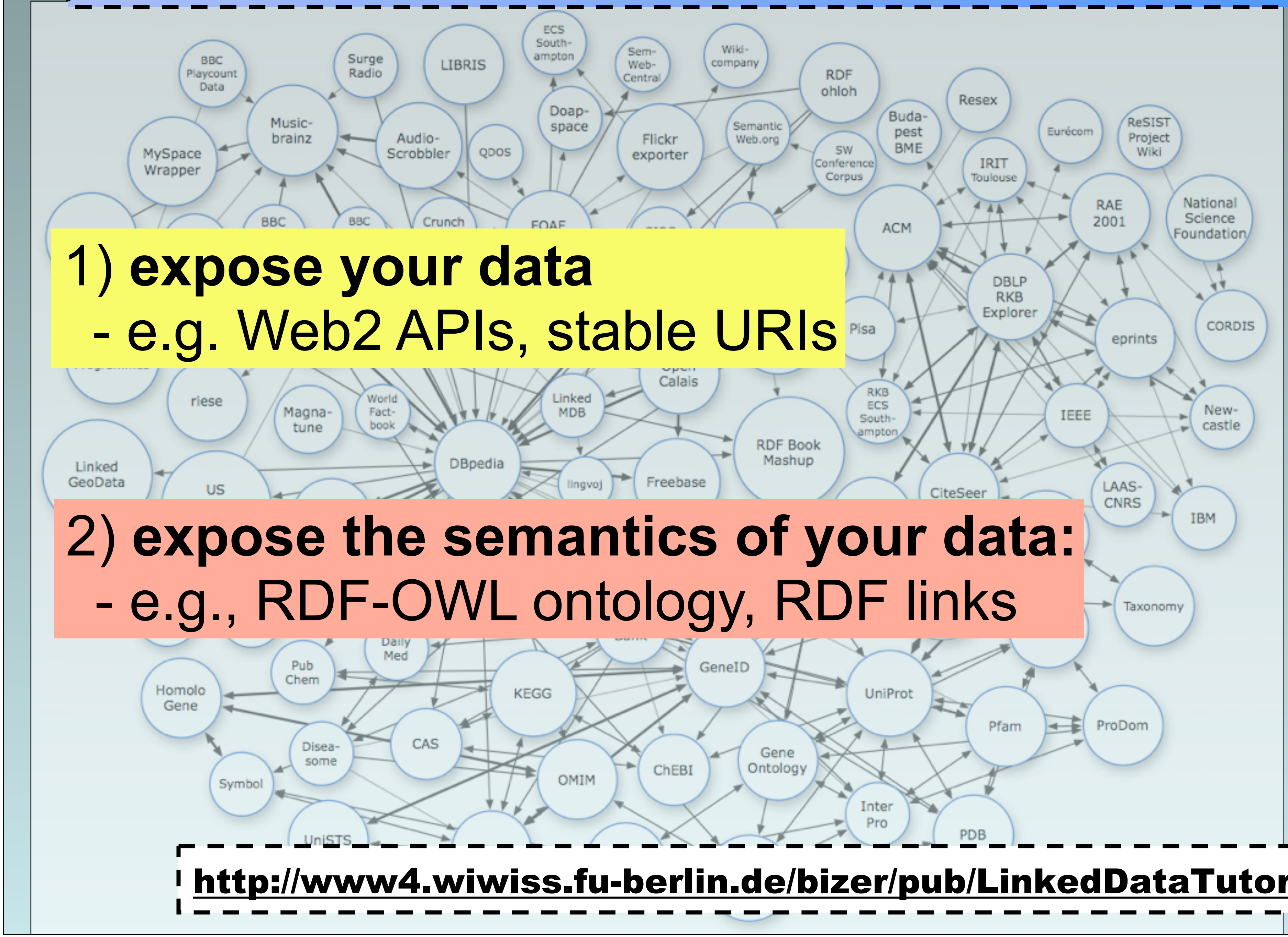
Approaches: the Linked-Data initiative (II)



March 2009

http://linkeddata.org/

Two Essential Principles



1) **expose your data**
- e.g. Web2 APIs, stable URIs

2) **expose the semantics of your data:**
- e.g., RDF-OWL ontology, RDF links

http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

Potential pitfalls [1]: exposing the data model

- domain experts are needed, but are not enough

- the knowledge engineer must initially play a ‘socratic’ role (importance of stating the obvious)

- data schemas are not ontologies!

- Writing something in OWL does not make it an ontology! The key difference is not the language the **intended use**

- making representational choices at the ***highest*** **level of** ***abstraction***, **while still being as clear as possible about the meaning of terms**

- taming the ontological beast

- ‘shallow’ ontologies vs ‘deep’ ontologies e.g. Foaf vs Dolce

- reusability vs expressiveness

- create of reuse?

Potential pitfalls [2]: exposing the data

- DataBase or not DataBase?

- working with relational DBs in many cases stills seems to be the most effective solution

- eg performance, scalability, support in MySQL and PostgreSQL

- powerful webApp frameworks usually rely on RDBs (Django, Rails, Ignite)

- A crossroad, depending on whether: a) DB work has ended b) DB work is ongoing

a) we choose/create an ontology that satisfies us, export our DB in RDF and put it in a SPARQL-enabled triple store (= other people may access our data easily)

b) the triplestore needs to be constantly updated/monitored .. need to have a time-indexed triplestore!

Wrapping up..

- the LTB will provide a SPARQL endpoint
  - data + ontology
  - we’re aiming at Sept 2010
- REED is moving its first steps in this direction
  - might not be RDF though..
  - first we need a solid framework for managing ‘live’ data
- comments?

thanks....