

Contested relationships and RDF

Jeff Good, University at Buffalo (jcgood@buffalo.edu)

Humanities Data: Tools for Annotation and Access
Digital Humanities Initiative at Buffalo

Overview

- A conceptual distinction with consequences for database construction
- Discussion of use of an interesting emerging technology (RDF) as part of the solution for modeling that distinction

Two kinds of databases

- **Reference database:** A database containing data whose structure is presumed to be well understood used to facilitate access to known information
- **Research database:** A database containing data whose structure is poorly understood used to facilitate research on the nature of the knowledge domain from which the data is drawn

Database examples

- An online dictionary, like the OED, is a reference database
- The British National Corpus is a reference database
- Lexicons created by field linguists describing new languages usually start out as research databases...
- ...but ultimately become reference databases

Ignorance

headword

POS

gloss

puella n girl

Database structure

Example data

Confusion

headword

POS

gloss

genitive

dative

accusative

Database structure

puella n girl

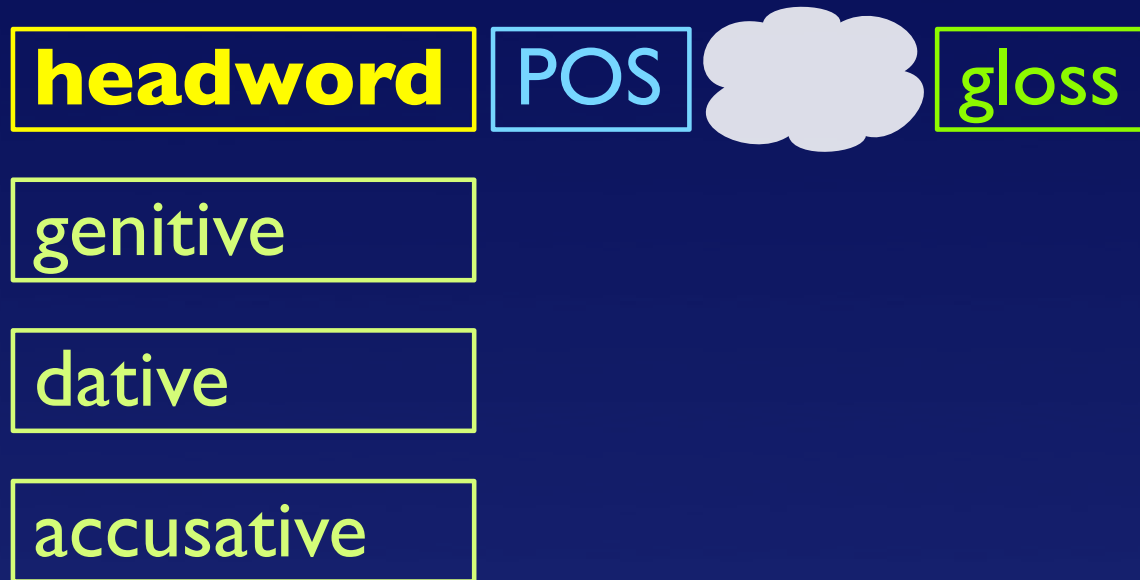
puellae

puellae

puellam

Example data

Enlightenment



Database structure

Example data for the word 'puella'. The word 'puella' is in a yellow box, followed by a blue 'n', a white cloud icon, and a green 'girl'. Below this are three lines of text: 'puellae', 'puellae', and 'puellam'.

Example data

Codification

headword

POS

DEC

gloss

genitive

dative

accusative

puella n 1st girl

puellae

puellae

puellam

Database structure

Example data

Reconstruction

headword

POS

DEC

gloss

puella n

1st

girl

Database structure

Example data

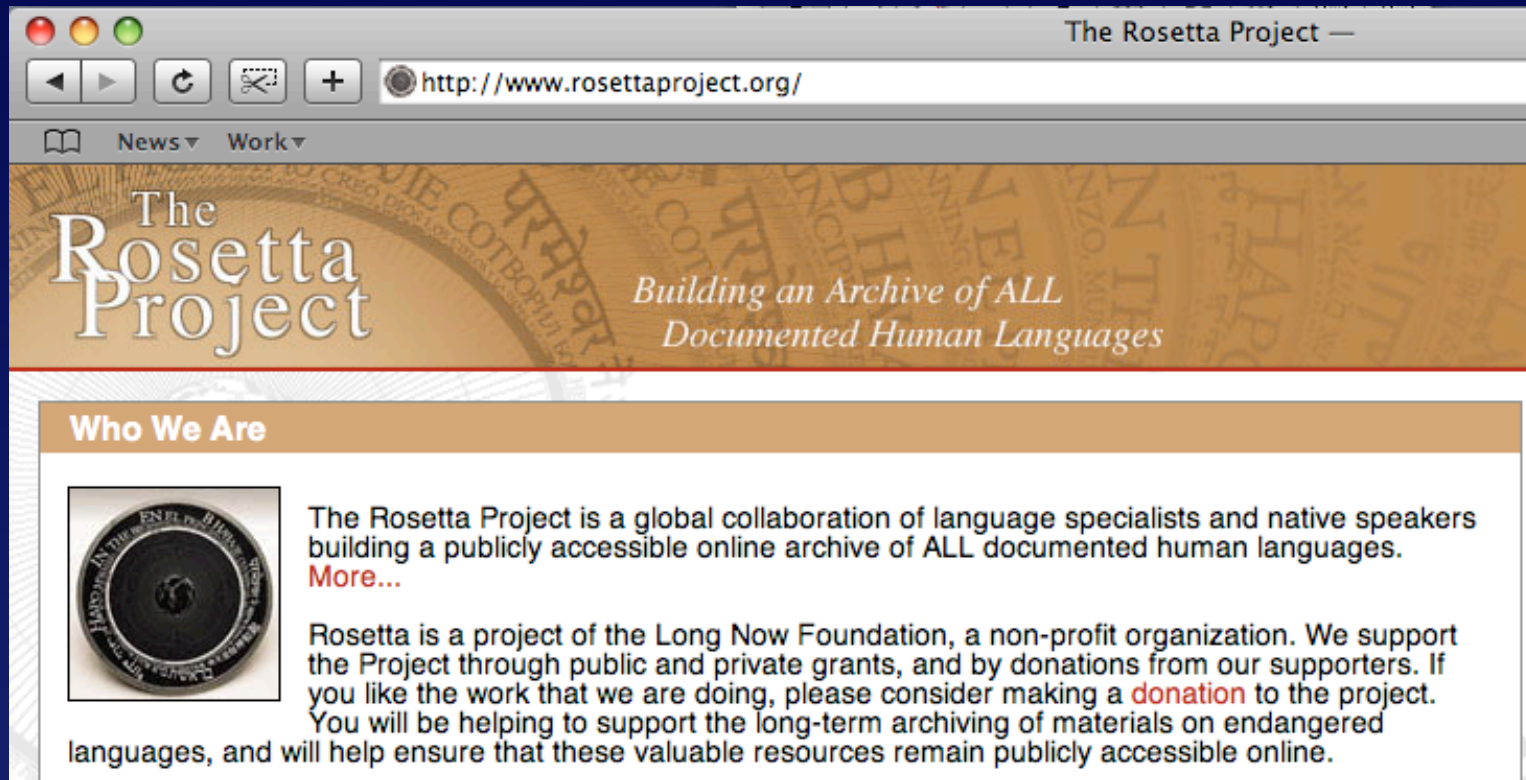
Tension

- Some stability in the database is necessary
- Too much stability hinders achieving real research results

The Rosetta Project



The Rosetta Project



- “Building an archive of ALL documented Human languages.”
- **Not** conceived of by a linguist.

Design problems

- There are many, many design problems: User interface, data modeling, long-term preservation...
- Data organization
 - Non-contested data
 - vs.
 - Contested data

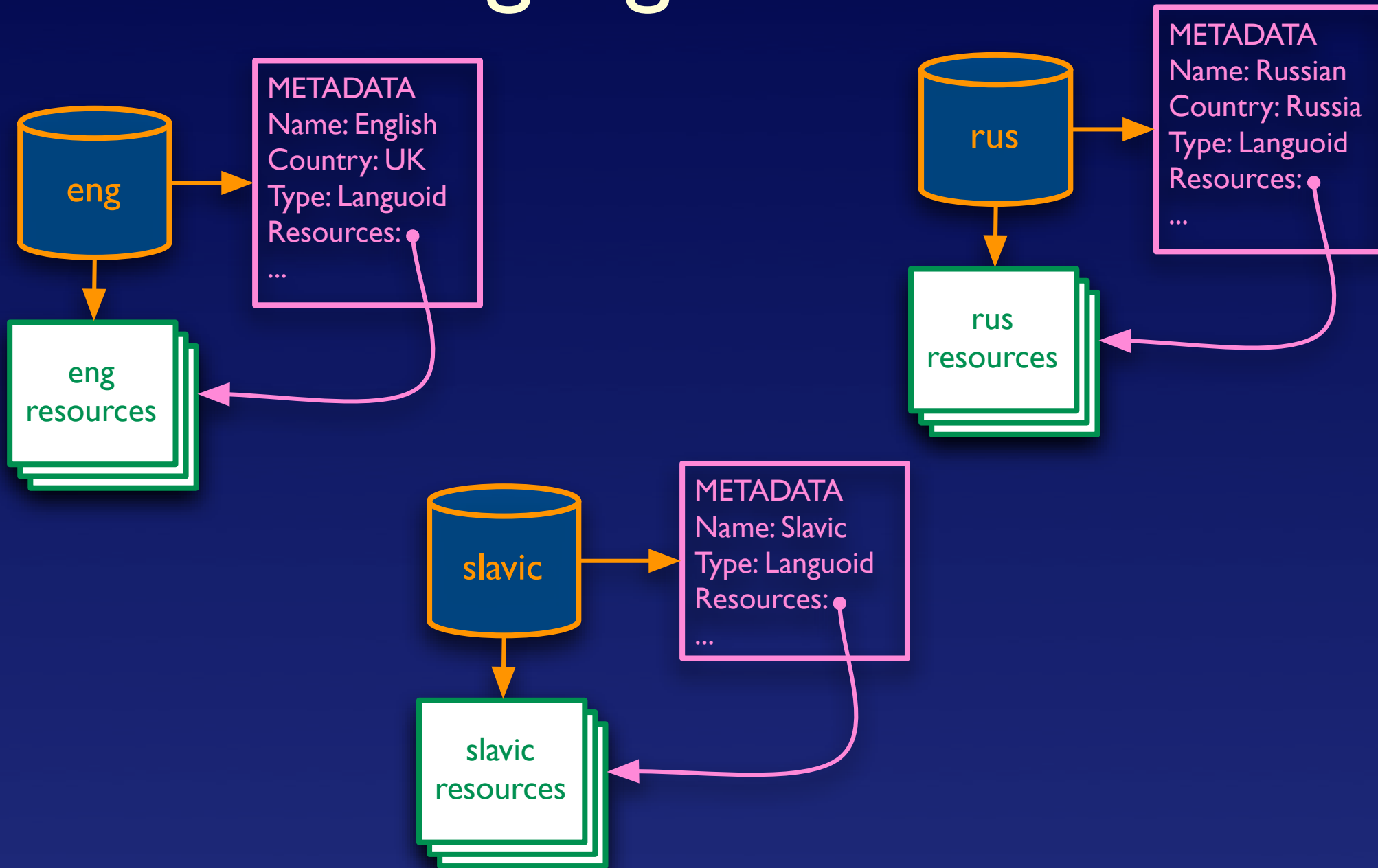
Modeling

- **Core problem:** How can we make use of contested data without “contaminating” the non-contested data?
- **Data to model:** Languages and language relationships

Data types

- **Non-contested**
 - Existence of a language (or language family)
 - Resource describes a given language
- **Contested**
 - Family tree for language
 - “Language”, “dialect”, “family”

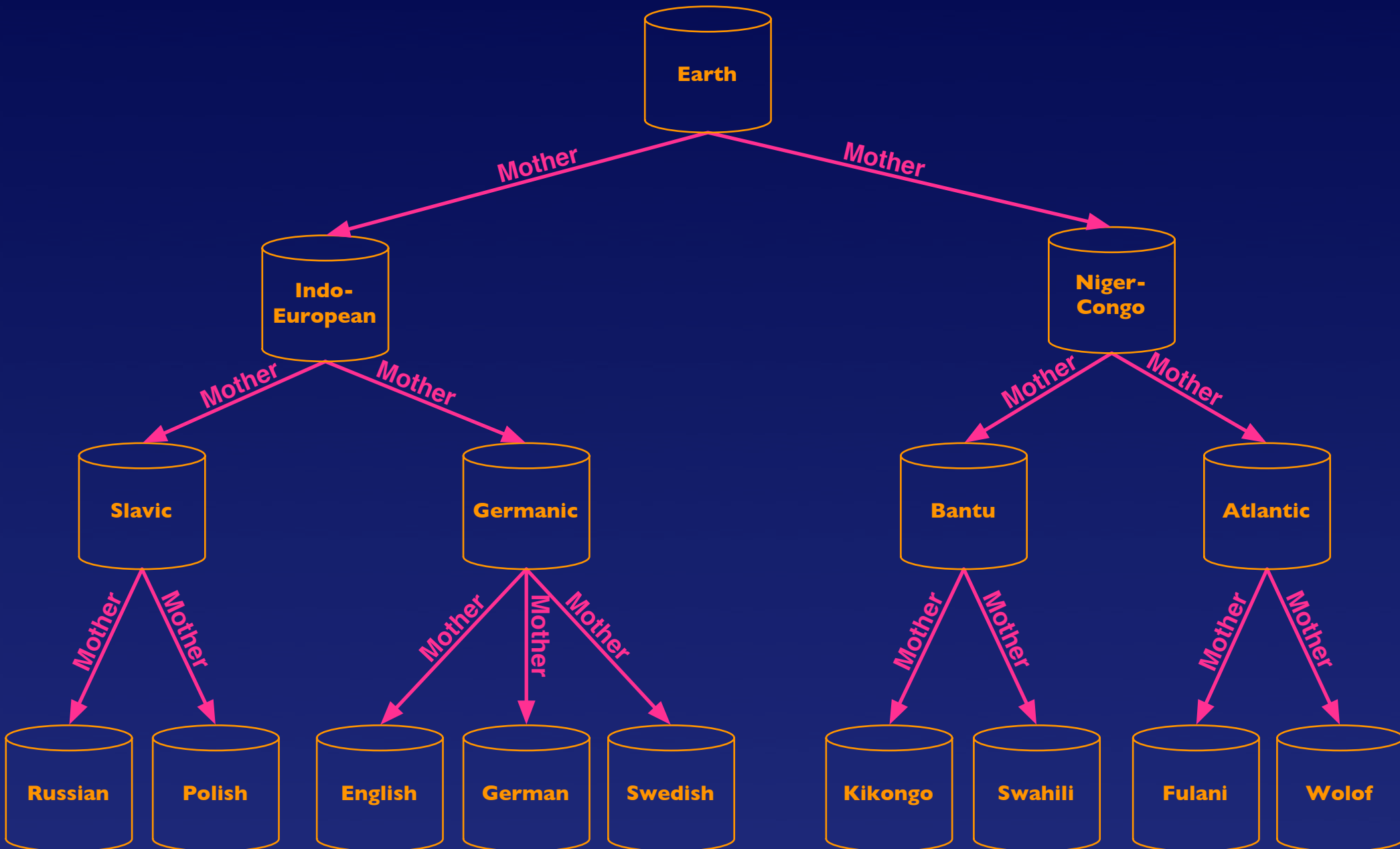
Language nodes



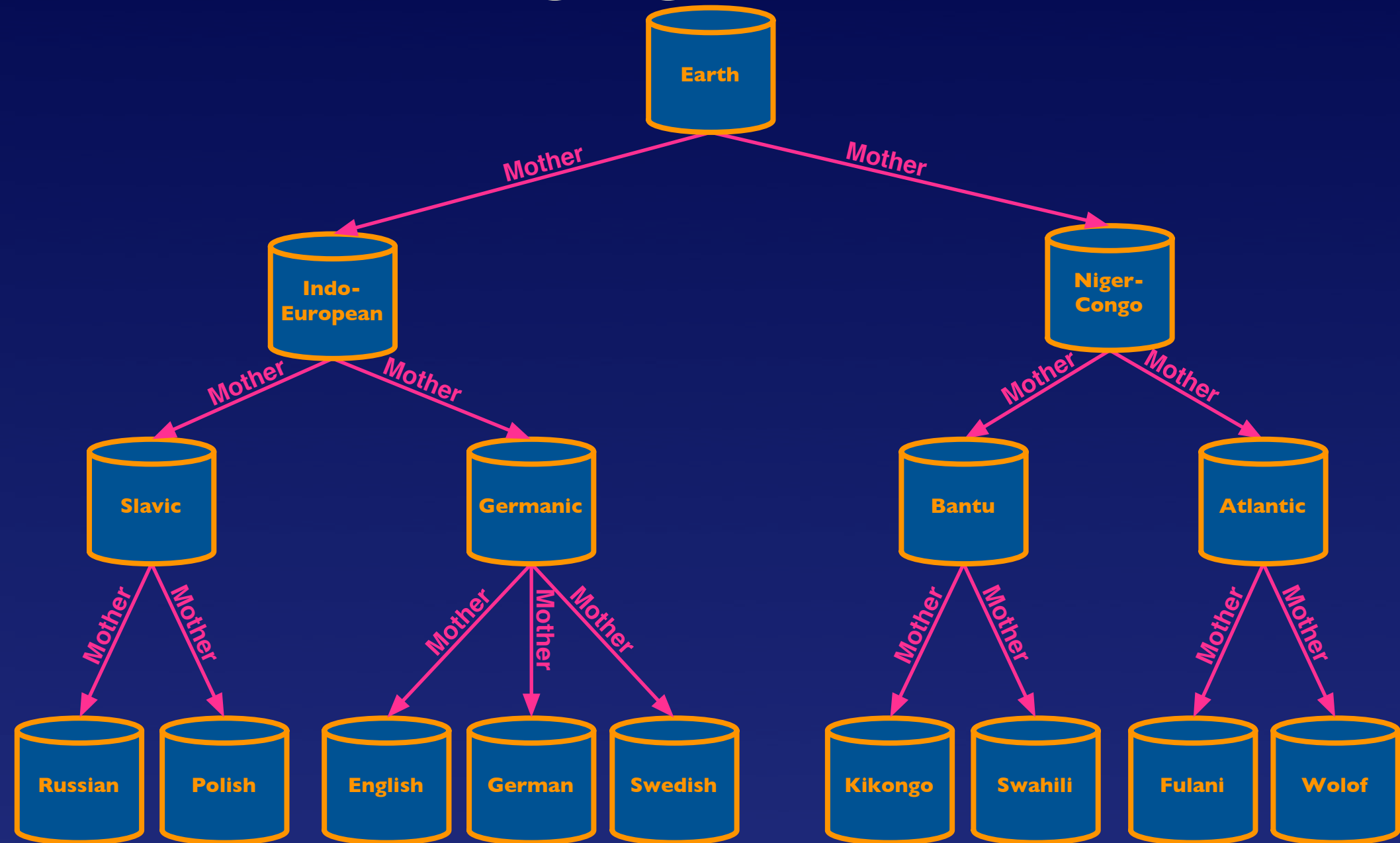
The node database



The relationship database



Merging the data



Implementation

- **Language nodes:** Object-oriented database
- **Relationships:** RDF database
- **RDF:** Resource Description Framework
 - Key component of the Semantic Web
 - Very good at modeling tree structures (among other things)

RDF: Leading ideas

- *Everything* should have a *universally* unique identifier
- Information can be expressed as a network of three-place statements consisting of a *subject*, *predicate*, and *object*



RDF: Example



RDF: Example

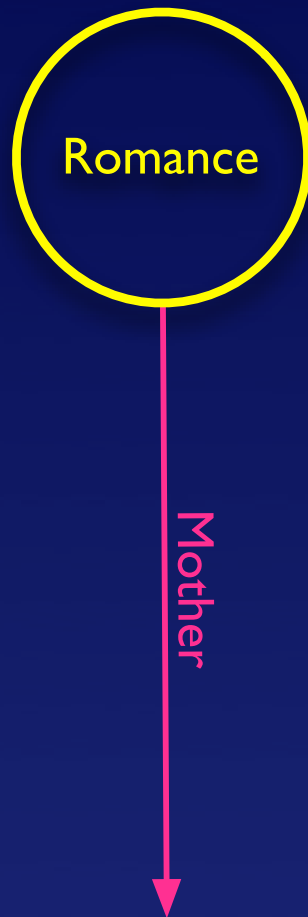


Romance



<http://rosettaproject.org/archive/Romance>

RDF: Example

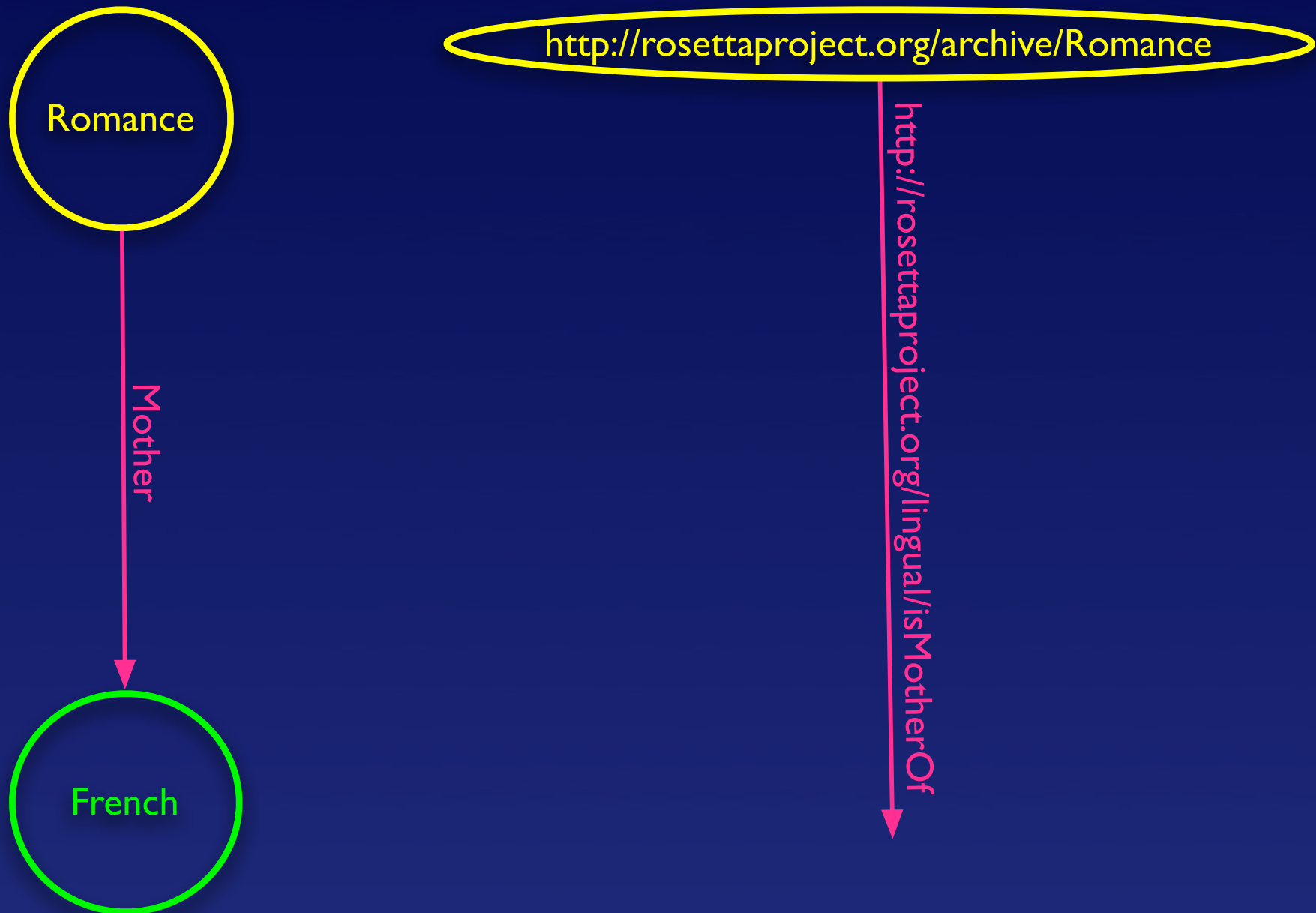


<http://rosettaproject.org/archive/Romance>

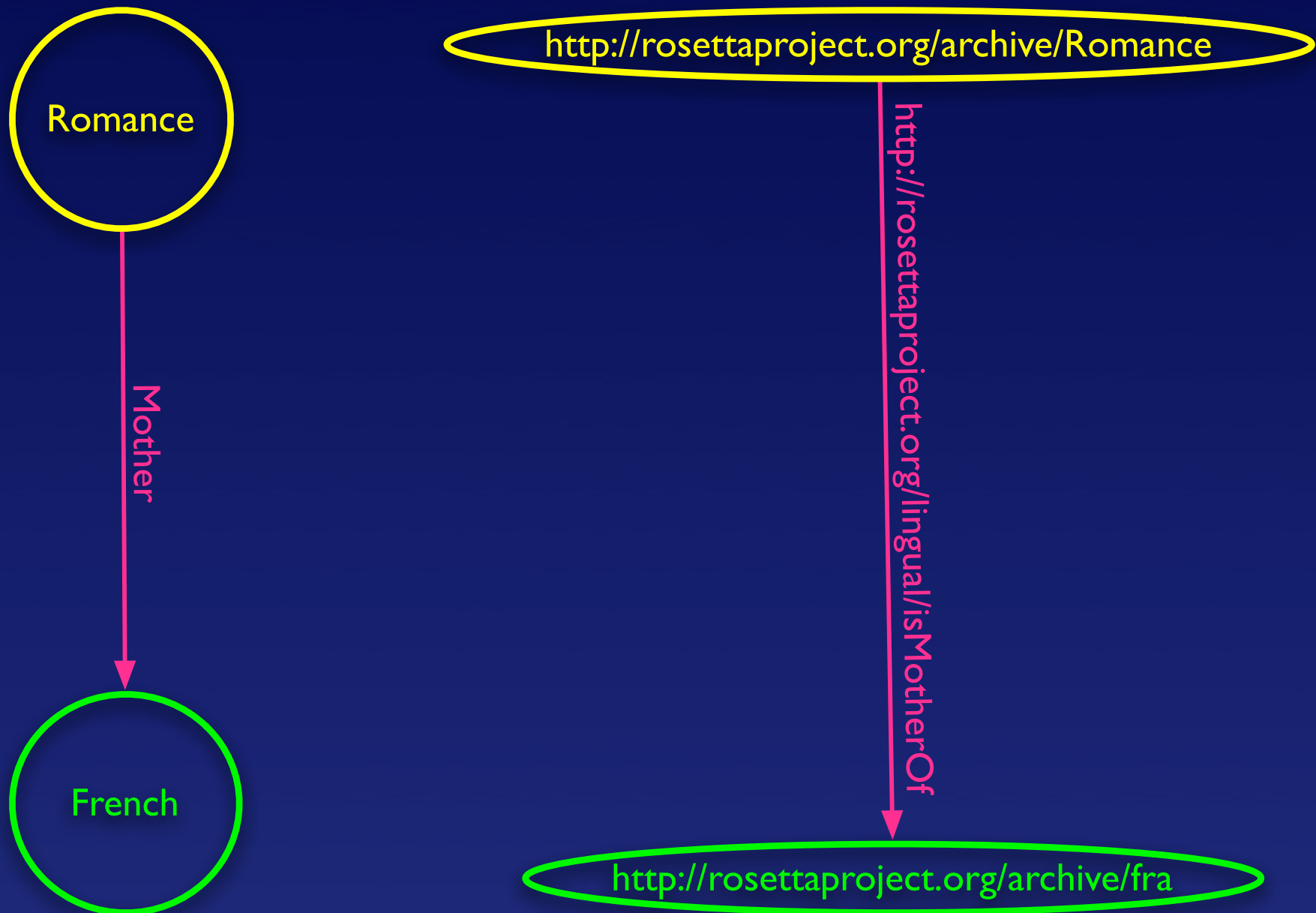
RDF: Example



RDF: Example

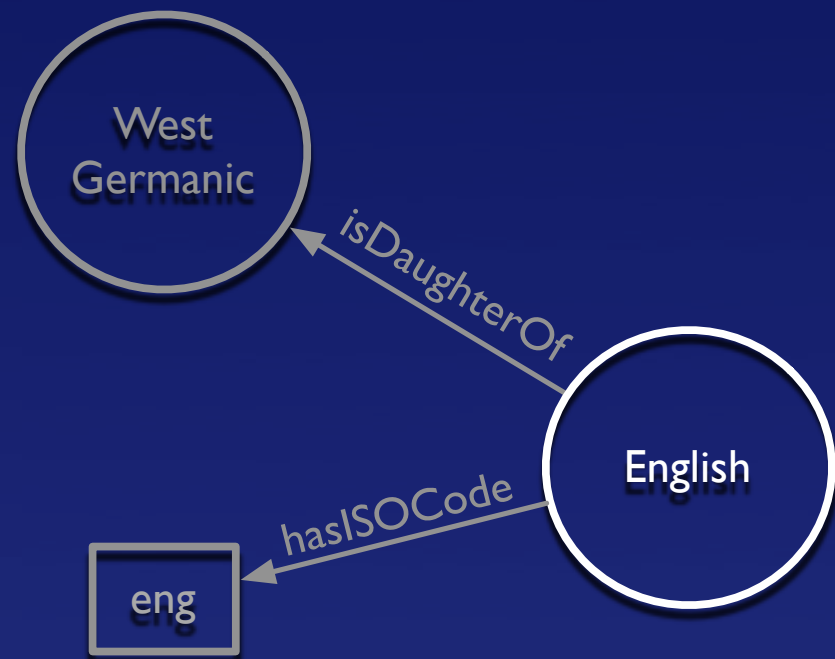


RDF: Example



Just what is RDF?

- RDF can be understood as a simplistic—and therefore “universal”—data model
- RDF can be expressed in various ways, including XML
- Unlike “classic” XML, it is not good at encoding the structure of documents
- It is good at expressing relationships among documents, annotations, and other kinds of “objects”



Quirk et al.
(1985)

containsExample

John saw Mary

Data

documents

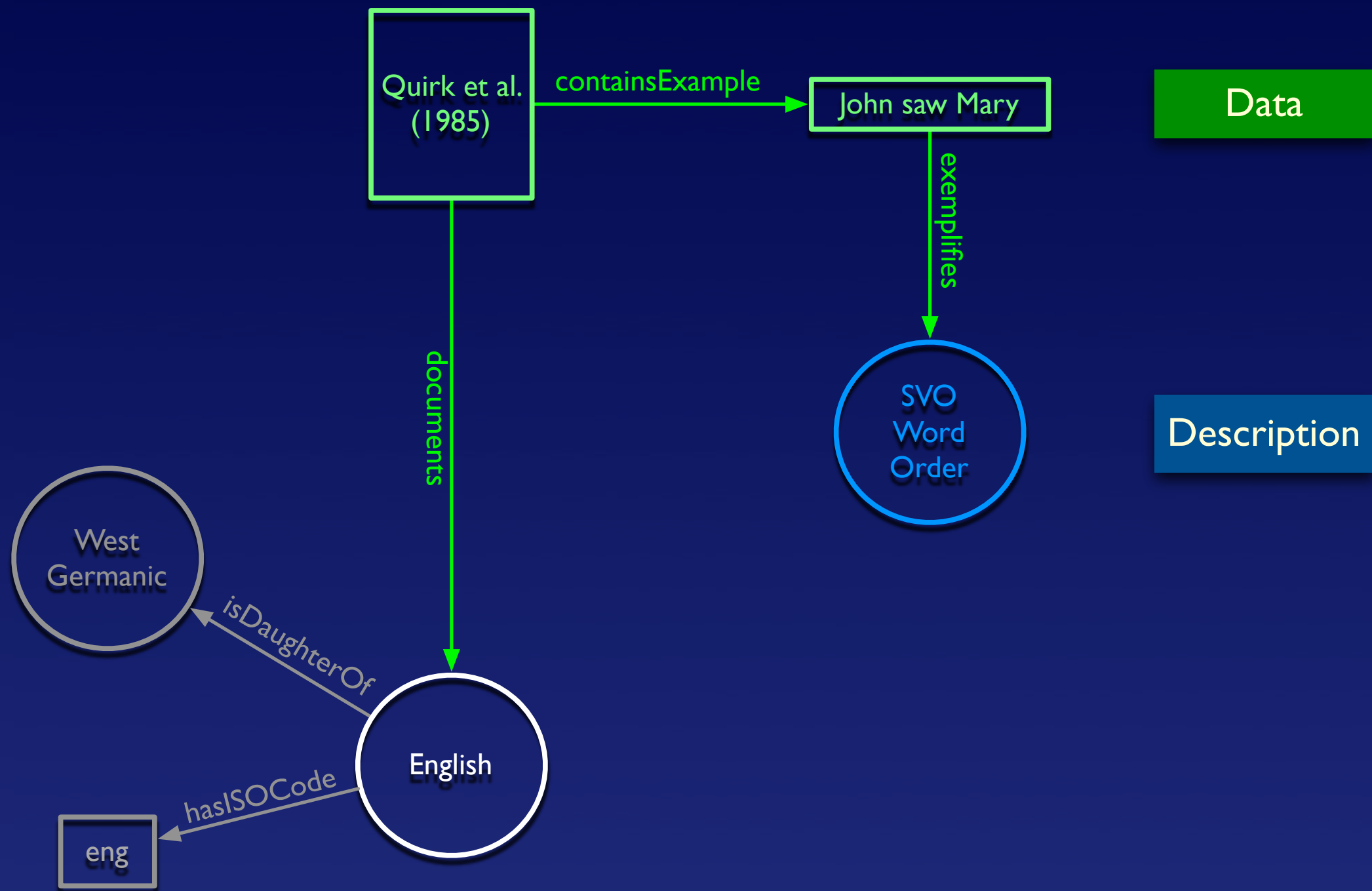
West
Germanic

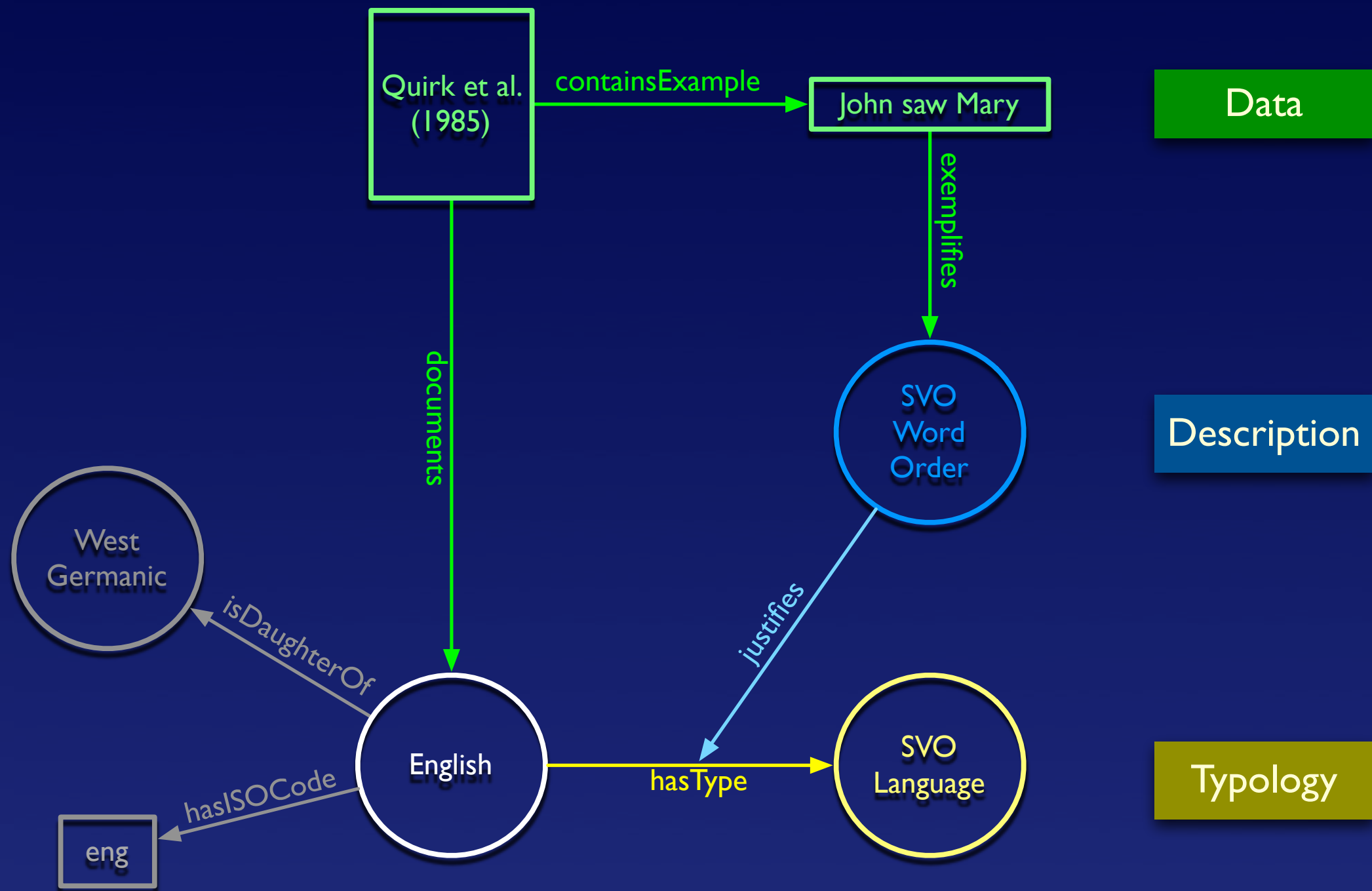
isDaughterOf

English

hasISOCODE

eng





RDF and relationships

- RDF provides a good way to model relationships
- Relationships are what is usually contested
- By keeping relationships separate from “objects”, we can use contested information without depending on it
- There are ways to do this not involving RDF: The same conceptual issue is usually amenable to multiple implementations