

The GOLD Community Vision

Scott Farrar

Transregional Collaborative
Research Center,
Universität Bremen



Goals of the Talk

- Describe a model for the GOLD Community of Practice
- Discuss the data and knowledge components of the model
- Focus on a Web implementation of the model
- Discuss the representation language for each component
- Set the stage for discussing services to be built around the model (talks by Lewis, Simons)

Some Special Terms

- *Web resource*: anything with a URI.
- *RDF*: A Web language for expressing relationships among resources.
- *OWL*: A quasi-standard Web ontology language that builds on RDF---captures knowledge about resources.
- *Web service*: server-side application that manipulates Web content for a client.

What is a Community of Practice?

In general:

- A group focused on a common activity or having a common sense of purpose
- A group that shares knowledge about a given domain

Specifically:

- A group of researchers consistently applying the same meaning for a given terminology
- A group sharing a common tool or data set

Examples of Communities of Practice

- Users of the IPA
- WALS contributors
- OLAC metadata providers
- DOBES sponsored field researchers

- Users of GOLD

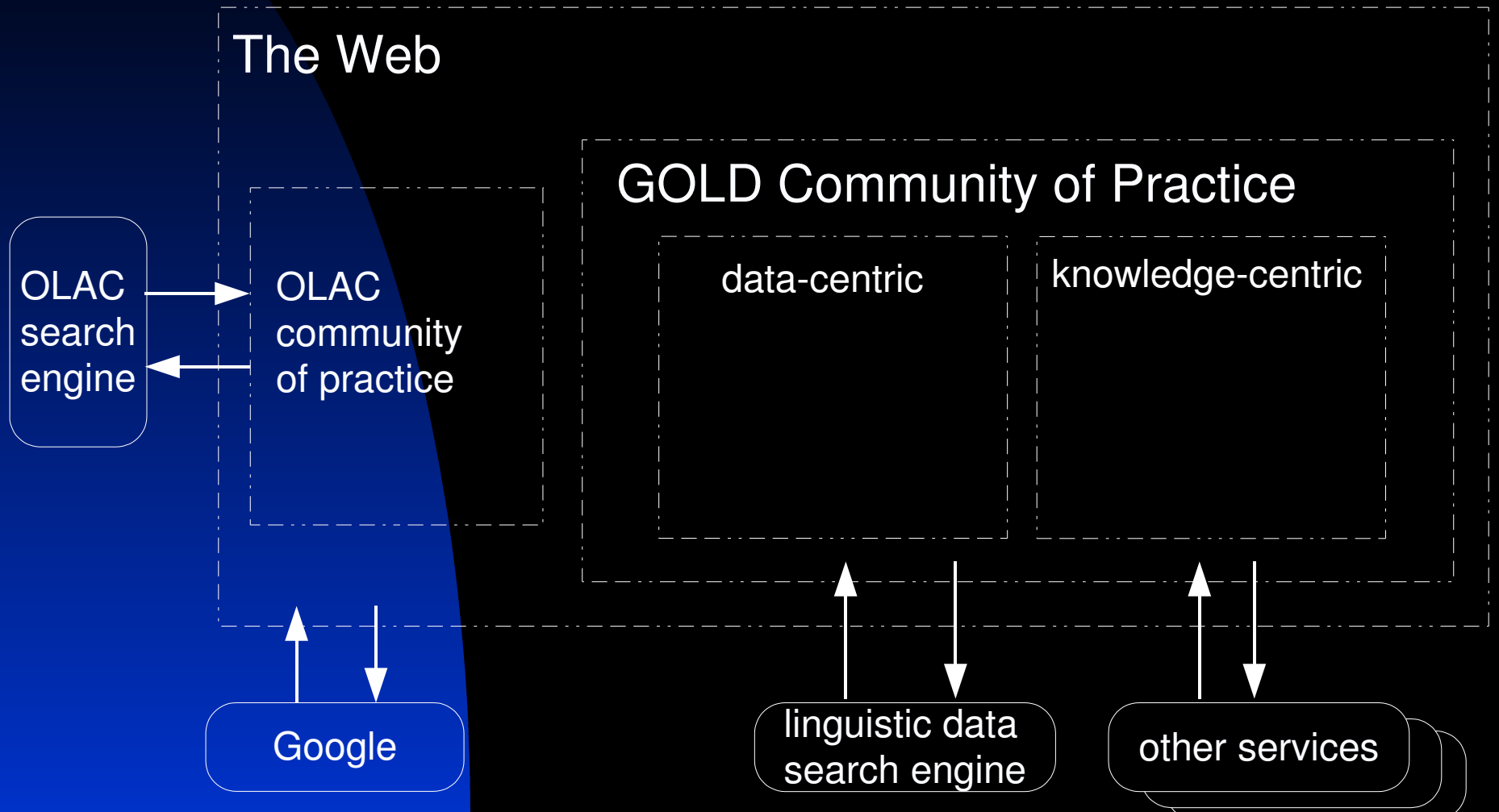
Guiding Principles

- Openness of Encoding and Markup
- Explicit definition of terminology
- Use of open source (no proprietary tools with secret or unpublished formats)
- Interoperability
- Open access (where possible)
- Broad community involvement
- Priority of data over knowledge

Why Establish a Community of Practice?

- Rapid access to data
- Verification of integrity of data
- Sharing code for building data creation tools (FIELD)
- Automated search over massive amounts of data (ODIN)
- Codification of the knowledge of linguistics (GOLD)

The Big Picture



Challenges to Building the Community of Practice

- Disparate data structures across resources
- Disparate markup used across resources
- Need to achieve interoperability without sacrificing local control over data resources.
- Need for (semi-)automation

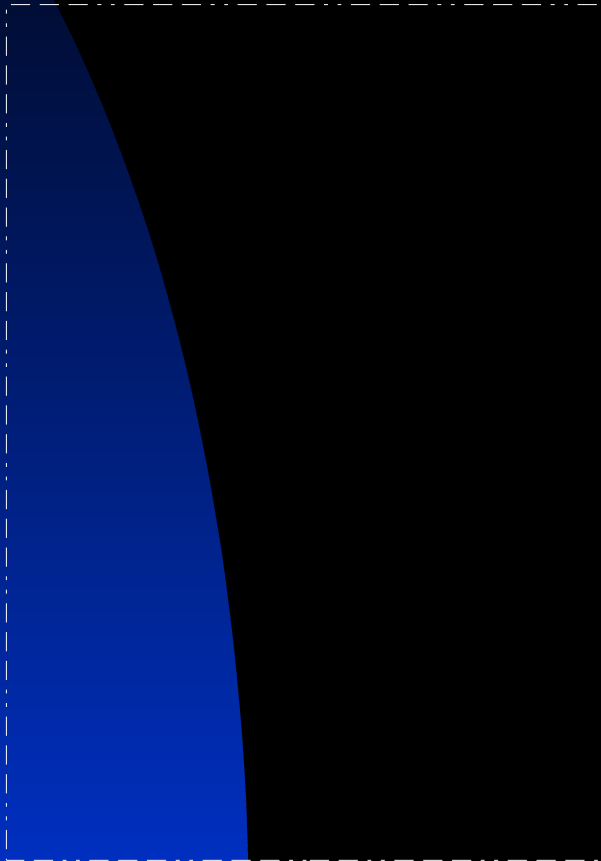
- It's difficult to establish trust within the community....that's why we're here!

Components of the GOLD Community of Practice

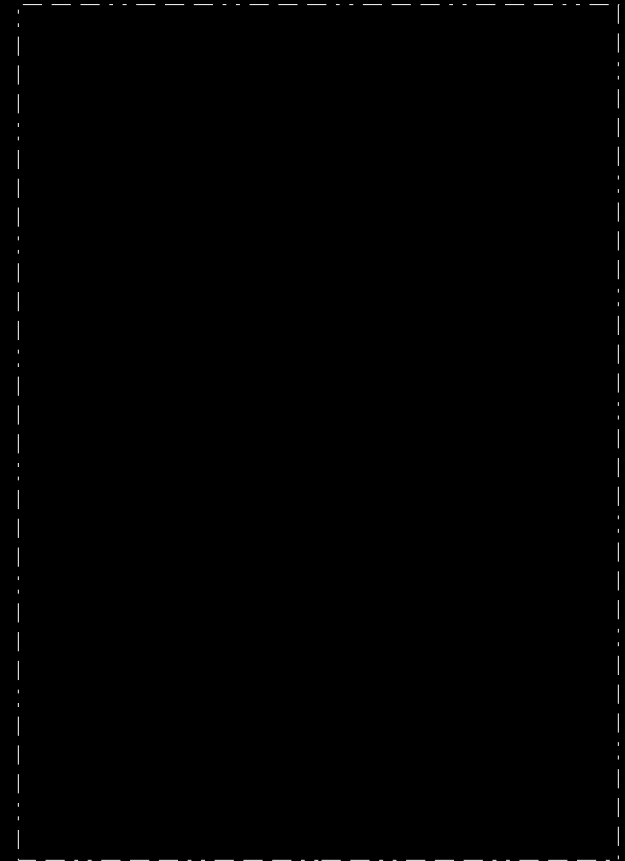
- Data-centric components
 - the DATA, DATA, and more DATA
 - descriptive resources about DATA (metadata, bibliographic,...)
 - terminologies
- Knowledge-centric components
 - knowledge about particular languages, theories, structures
 - general knowledge of linguistics (GOLD)
 - foundational knowledge (an upper ontology)

Components of the GOLD Community of Practice

data-centric



knowledge-centric



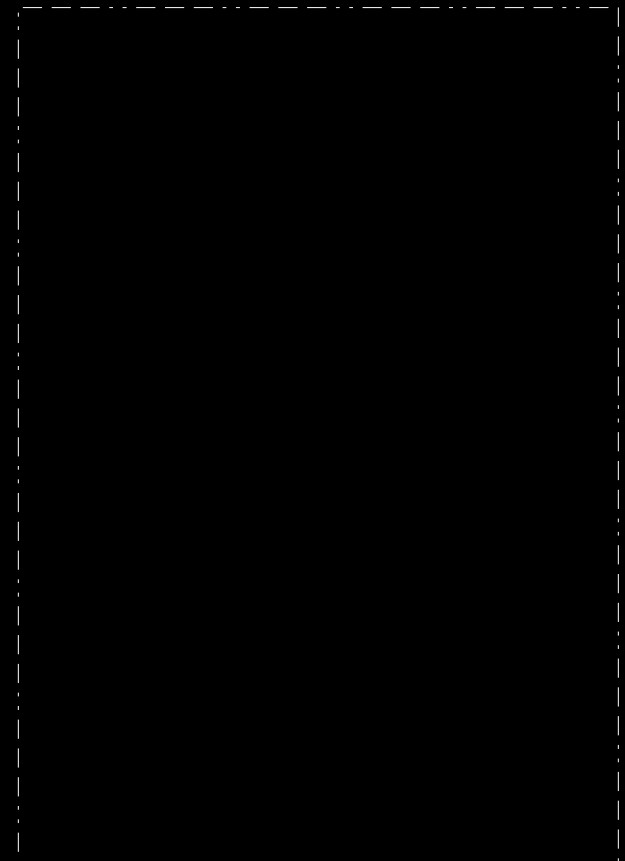
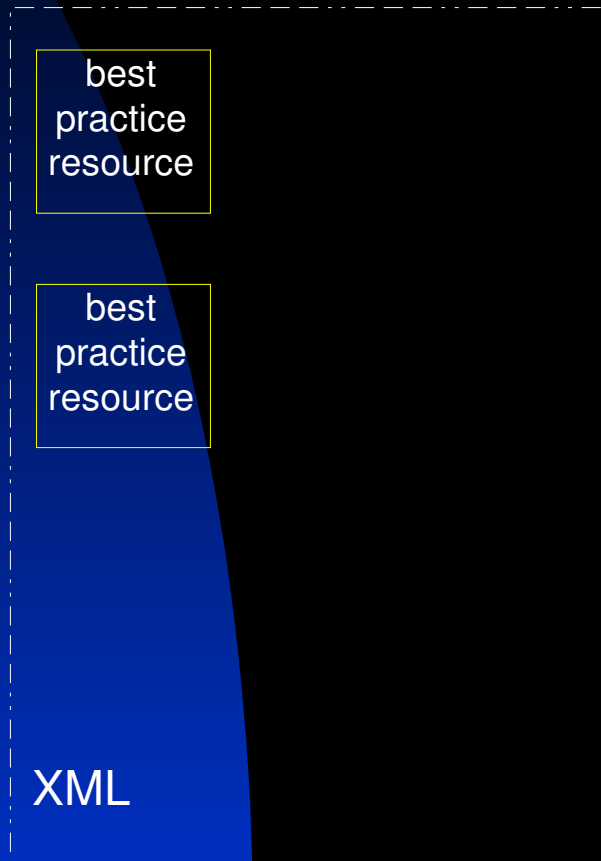
DATA: Best Practice Resources

- encoding: Unicode
- markup language: XML (with accompanying DTD/Schema)
- markup content: descriptive- vs. display-oriented
- Basically the suggestions of Bird and Simons (2003) *Language*, 79. and the E-MELD Project.

Components of the GOLD Community of Practice

data-centric

knowledge-centric

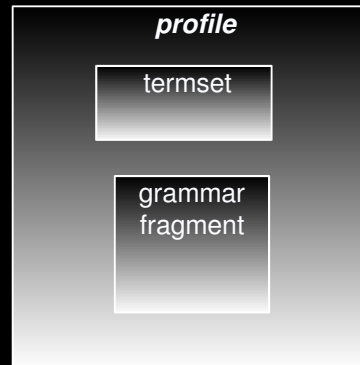


DATA: Best Practice Resources

- Problem: The markup in a data resource needs to be highly articulated to achieve any degree of interoperability (and automated migration).
- Solution: Construct a stand-off resource to clarify markup.
- Benefits: Data resource can be maintained locally, but can be migrated upwards in the model to inform the knowledge components.

DATA: Descriptive Profiles

- An XML document containing information about a best-practice data resource:
 - a Term Mapping
 - a Grammar Fragment



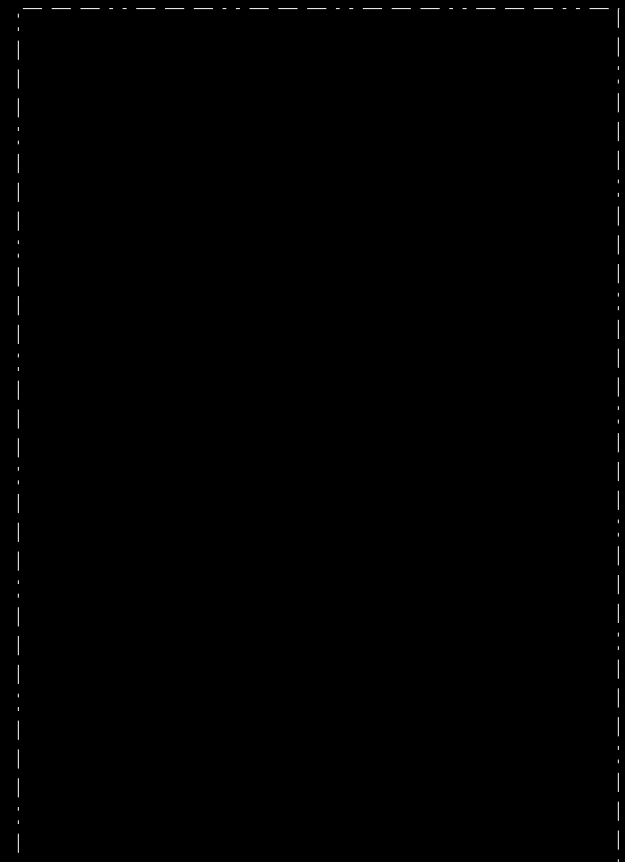
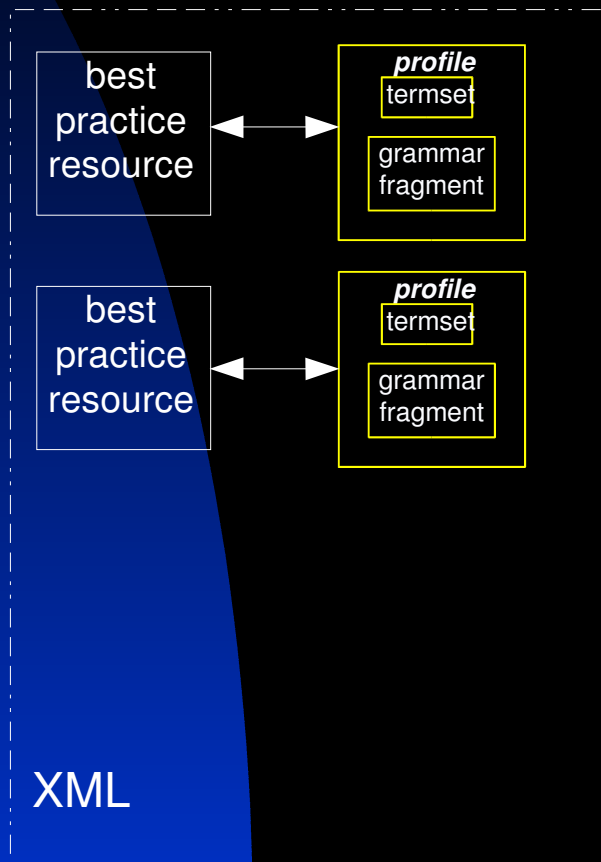
DATA: Descriptive Profiles

- Term Mapping:
 - A pair consisting of a markup element and an element in an ontology
 - A prose description of each element.
- Grammar Fragment:
 - A partial grammatical description of a resource, a phoneme inventory, list of tenses, some syntactic pattern expressed in a recognized data type (e.g., feature structures) (see E-MELD, TEI, ISO)

Components of the GOLD Community of Practice

data-centric

knowledge-centric



DATA: Legacy Resources

- Problem: Most data on the current Web are not in a best-practice format—legacy resources.
- HTML, PDF, MSWord, misc. Web db's
- Shoebox, text files (better practice)
- Scholarly papers are full of linguistic data.
- Such legacy resources are increasing rapidly.
- So, the Web is a GOLD mine of data.

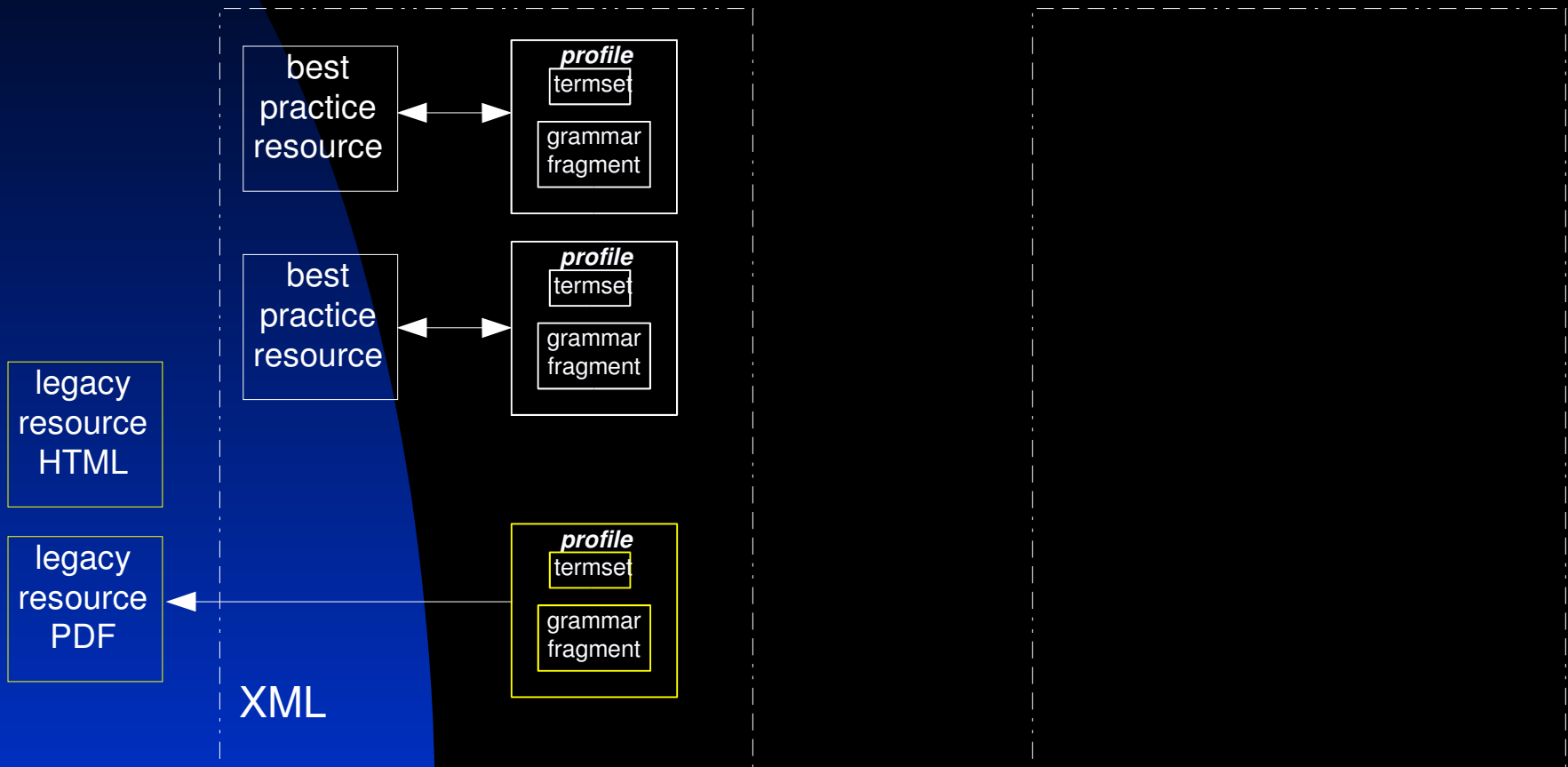
DATA: Legacy Resources

- Solution: Migrate whole resource to a best-practice format (labor-intensive).
- Or capture partial knowledge of legacy resource in a descriptive profile (more realistic).
- Benefits: A treatment of legacy resources draws on existing Web content. It's for free. Ensures success of the model by providing structured access to semi-structured Web content.

Components of the GOLD Community of Practice

data-centric

knowledge-centric



...Taking Stock

- Rich data environment in place.
- Locally maintained
- Potential for sharing resources (profiles, termsets)
- Best-practice requirements are satisfied

But...

- No real interoperability
- Data is only semi-structured due to inherent limitations of XML
- Much knowledge is implicit

Towards a Dynamic Knowledge Store (a Semantic Web)

- The implicit and explicit knowledge captured by the DATA can be abstracted to build a large KNOWLEDGE store on the Web.
- Such a resource can be the basis of many useful Web services.
- Broad interoperability is a real challenge.
- Whereas the model should ideally be bottom-up, a certain degree of top-down knowledge engineering is necessary.

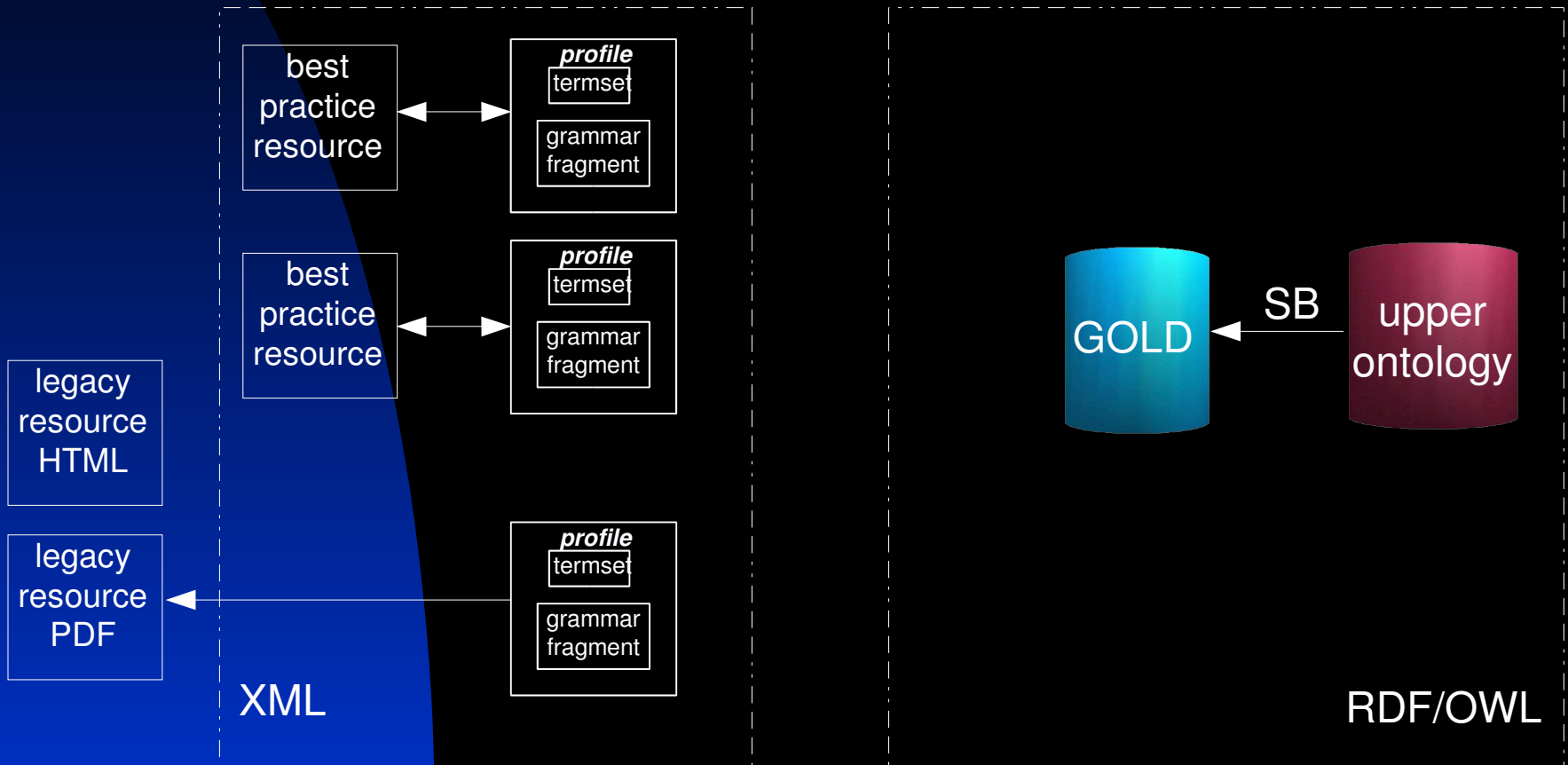
KNOWLEDGE: GOLD

- Problems:
 - Community acceptance is difficult to establish
 - Ontological modeling is hard (correct breadth and depth)
- Solutions:
 - Community involvement (Oversight board)
 - Use tools of formal ontology
- Benefits:
 - Precise definitions (in form of rich axiomatization)
 - Codification of basic linguistic concepts
 - Relation to other fields

Components of the GOLD Community of Practice

data-centric

knowledge-centric



Problem: General vs. Language-Specific Knowledge

- General
 - “A verb is a part of speech.”
 - “A verb can assign case.”
 - “Gender *can* be semantically grounded.”
 - “Linguistic expressions realize morphemes.”
- Specific
 - “Bantu languages have noun classifiers.”
 - “Mandarin Chinese has an aspect system.”
 - “German has three genders.”

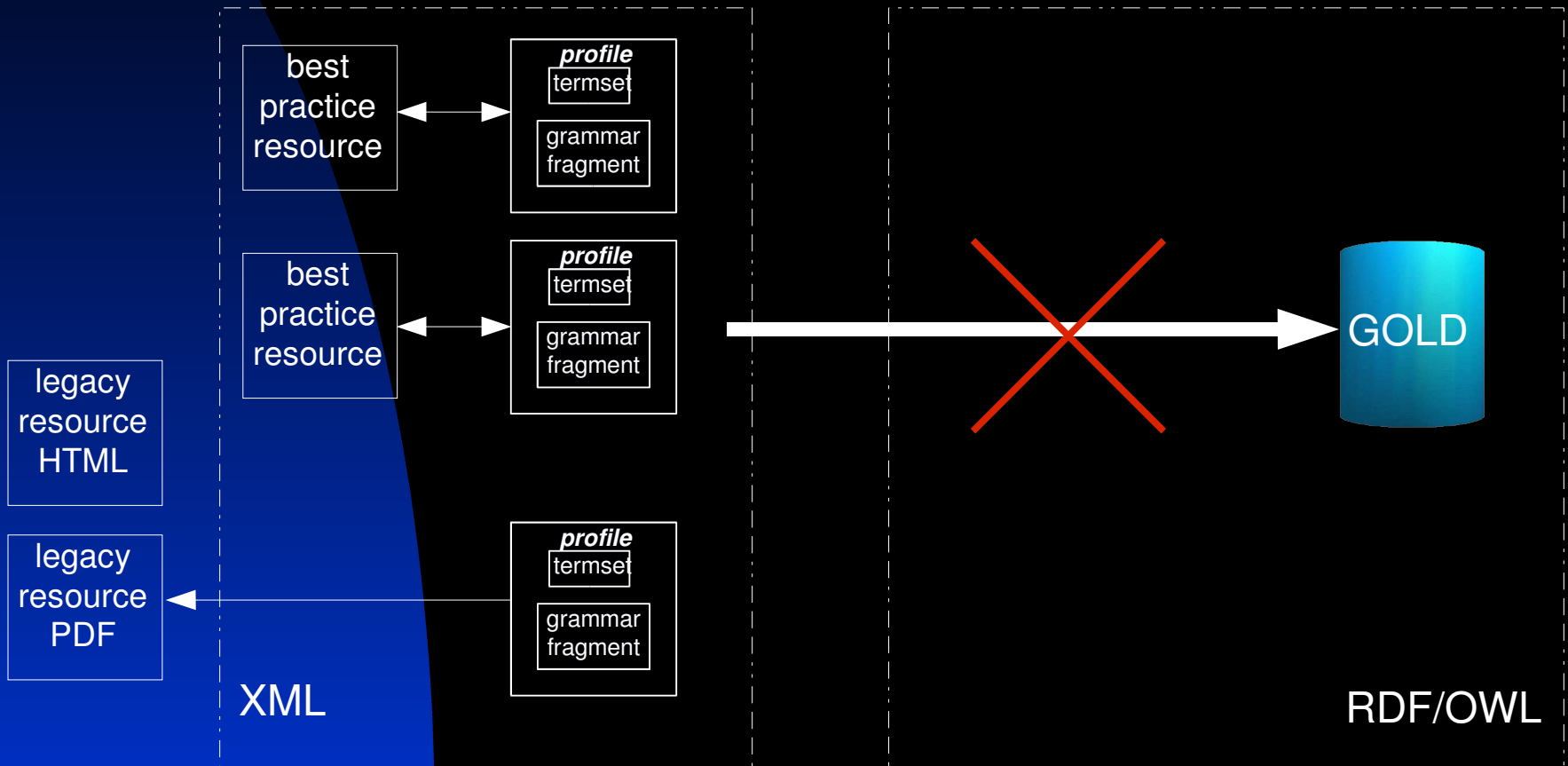
Problem:

**Linguists Don't Agree about
Linguistics!**

Components of the GOLD Community of Practice

data-centric

knowledge-centric



KNOWLEDGE: Community of Practice Extensions (COPEs)

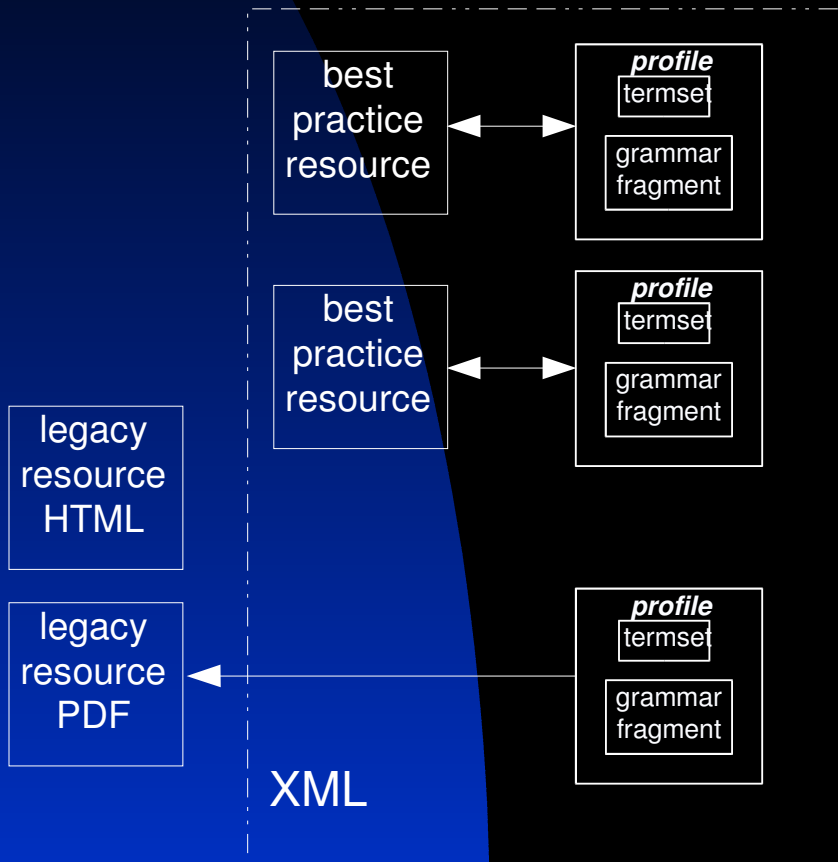
- Solution:
 - Reserve only the most fundamental knowledge of linguistics for the core ontology.
 - Create an ontological framework with GOLD at the center, but with the possibility of building community of practice extensions (COPEs).
- Dimensions of a COPE: level of analysis, theoretical perspective, language group, data type

KNOWLEDGE: Community of Practice Extensions (COPEs)

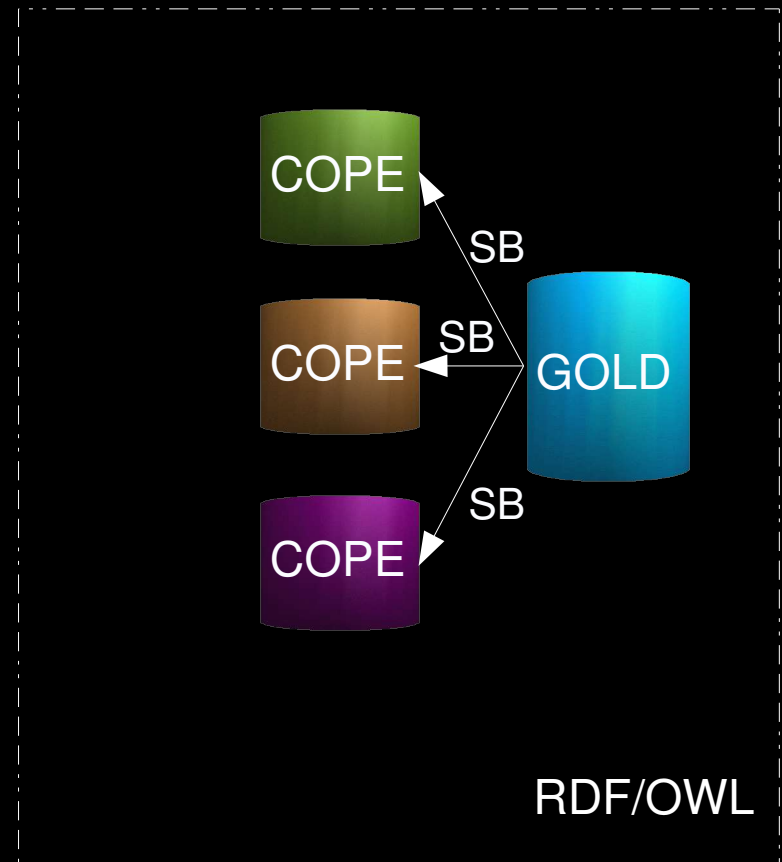
- Benefits:
 - Sub-communities can be individually maintained.
 - One change doesn't wreck the entire system.
 - Conflicting knowledge can be managed.
 - In general software is kept modular.

Components of the GOLD Community of Practice

data-centric



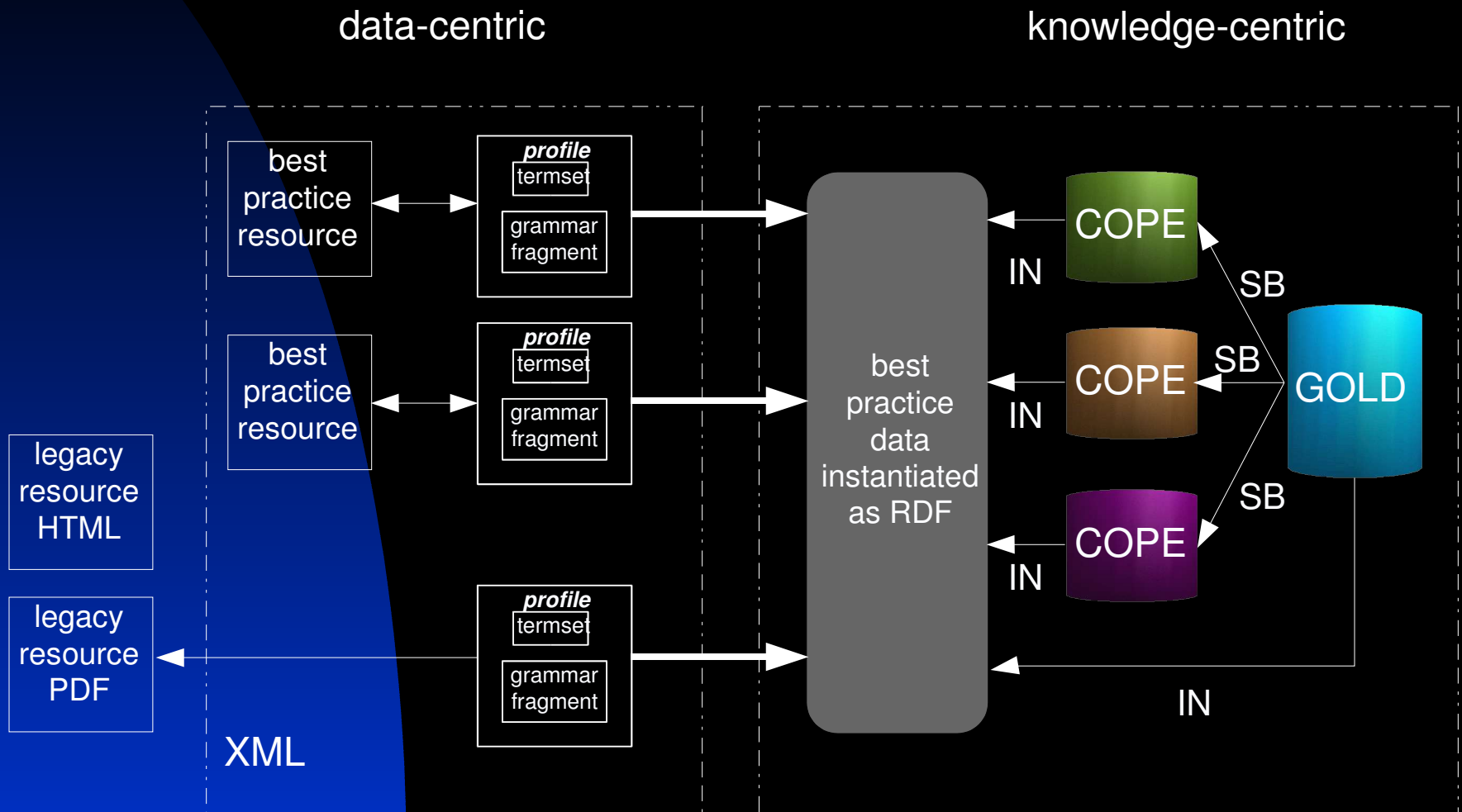
knowledge-centric



from DATA to KNOWLEDGE...

- The explicit and implicit knowledge of disparate best-practice resources can be migrated to a common, interoperable knowledge store.
- The data itself can be mapped to the knowledge store as instances of data types (e.g., a lexical entry, an occurrence of IGT).
- More generally, descriptive profiles contain information that can be mapped to instances of GOLD classes.

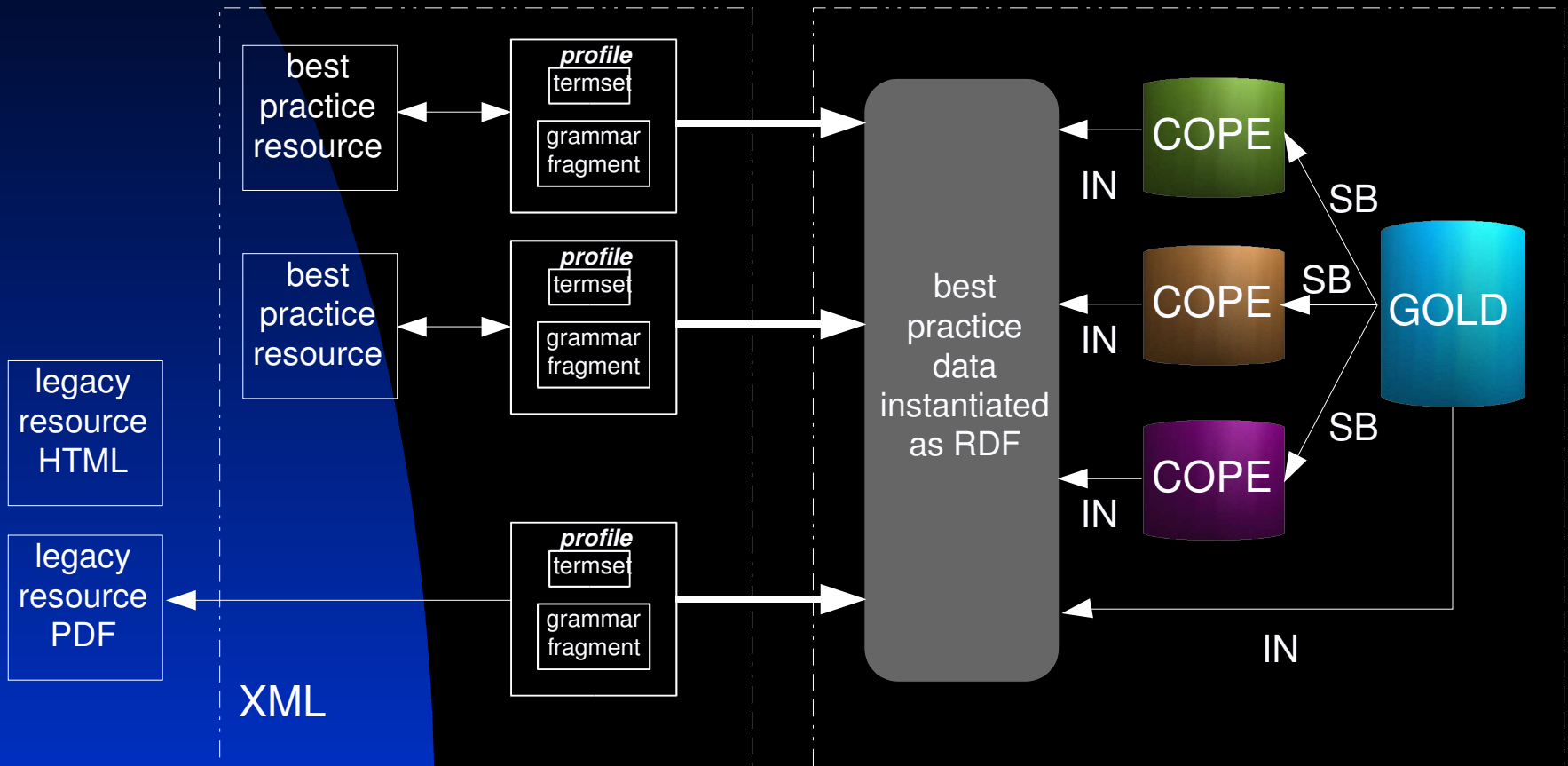
Components of the GOLD Community of Practice



Components of the GOLD Community of Practice

data-centric

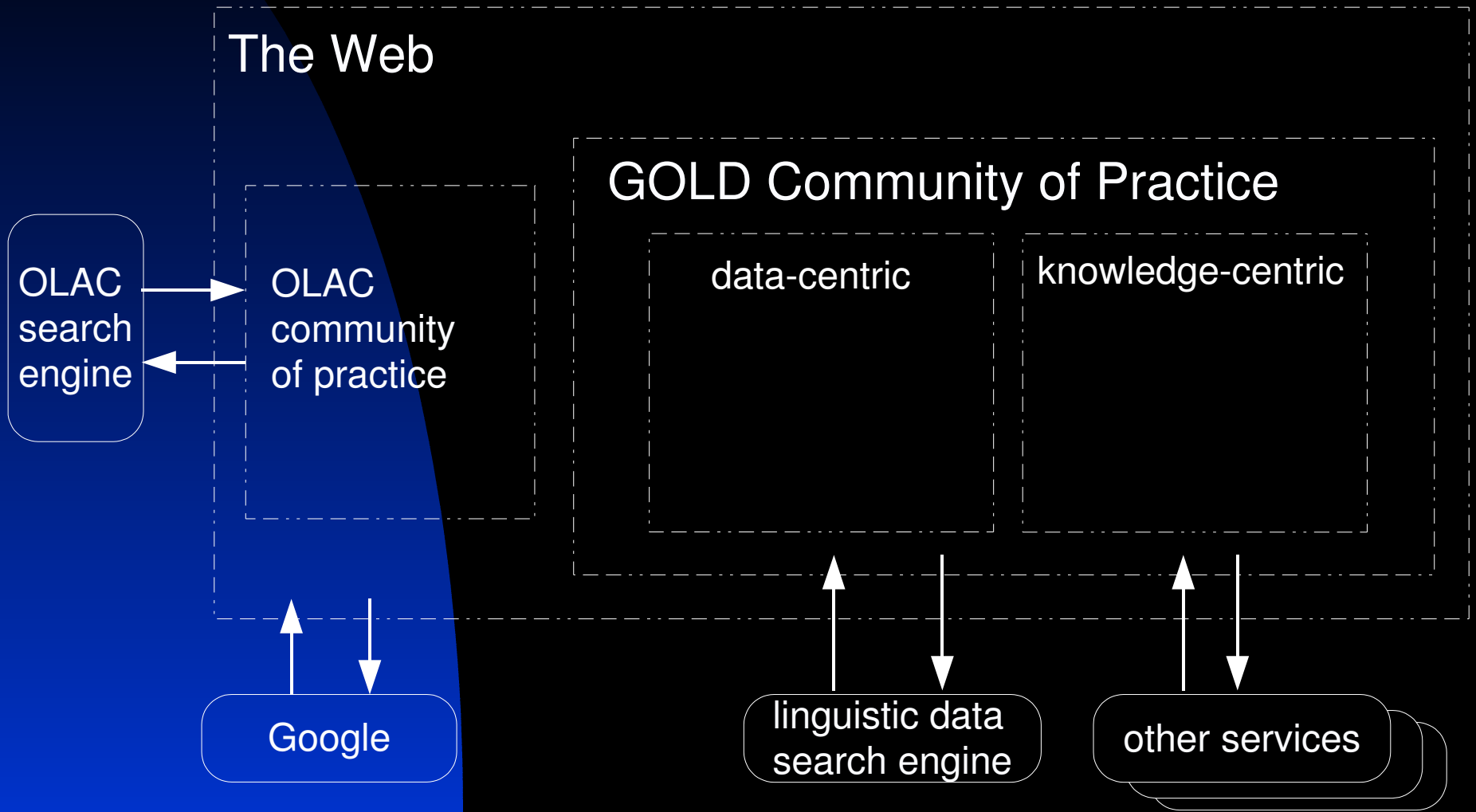
knowledge-centric



SERVICES

- Mapping best-practice resources to knowledge store is a service.
- Other examples:
 - tools to create best-practice and profile resources
 - tools to convert legacy to best-practice
 - search engine over the knowledge store
 - migration of portions of the knowledge store to optimized database systems
 - smart search with automated inferencing ability

Summary



Contact Info

- Contact: farrar@informatik.uni-bremen.de
- Website: <http://www.linguistics-ontology.org/>
- Full paper: (see workshop notebook)