Doing linguistics in the 21st century

Interoperation and the quest for the global riches of knowledge

(Full text)

Gary F. Simons

SIL International and GIAL

13 July 2007

Toward the Interoperability of Language Resources

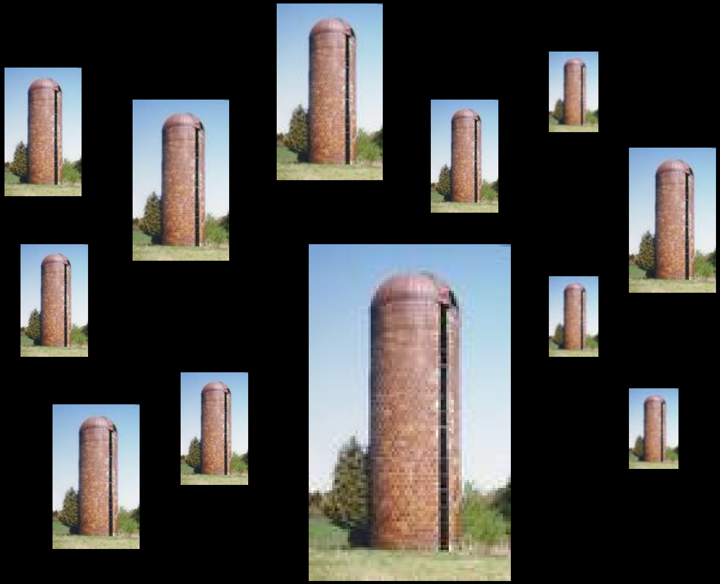
Stanford University, Palo Alto, CA

The quest for global riches

- Doing linguistics can be likened to a quest for riches
  - the riches of knowledge about language in general
  - the riches of knowledge about thousands of languages in particular

A possible future

20th Century



- Countless closed, local silos
- Economy of scarcity

21st Century

- One open, global treasury
- Economy of abundance

Scarcity vs. Abundance

physical world	digital world
finite shelf space	infinite shelf space
significant cost to produce and carry inventory	little cost to reproduce and carry inventory
limited capacity to carry inventory	virtually limitless capacity to carry inventory
publisher controlled content	user generated content
limited choice	unlimited choice
suboptimal matching of supply with demand	optimal matching of supply with demand

The dark side of abundance

- Problem
  - Unlimited choice = Overwhelming confusion
- Solution
  - Adopting constraints as a community that will limit our choices so as to promote interoperation within the community
  - Definition
    - Interoperability is the ability for two or more systems to exchange information or services and to make satisfactory use of what is exchanged.

Overview

- Adopting constraints as a community that will limit our choices …
  - An illustration from every day life of how standards will help our community
- … so as to promote interoperation within the community.
  - A vision for interoperation within our community in the 21st century,
    - expressed as a 12-point prescription for a cyberinfrastructure for linguistics

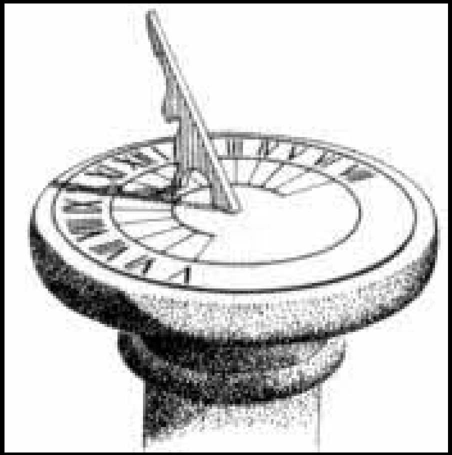
In defense of standards

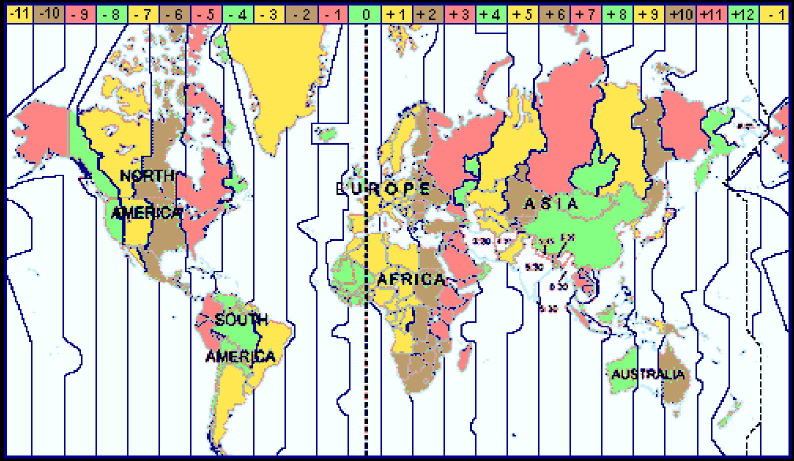
- Why are standards a hard sell for academics?
- Can we find metaphors that cast standards in a nurturant light rather than a strict one?
  - Linguistics as community
  - Development as freedom
- These ideas are developed in:
  - “Linguistics as a Community Activity: The Paradox of Freedom through Standards.” Forthcoming volume in honor of D. Terence Langendoen.

http://www.sil.org/~simonsg/preprint/standards.pdf

An analogy from every day life

- Two approaches to reckoning time
  - Solar noon vs. Standard time





- **Source:**
  - Blaise, Clark. 2000. *Time lord: Sir Sandford Fleming and the creation of standard time.* New York: Pantheon Books

Two approaches to local time

	Solar Noon	Time Zones
How it works	Local time is a continuous function; unlimited number of noons	Local time is a step function; exactly 24 noons
How it interoperates	Any east-west travel requires conversion by a number of min-utes to be looked up	Convert by even hours only over long distances
Optimized for	Convergence with reality (*i.e.*, truth)	Interoperation

The challenge for linguists

- In order to achieve interoperation, we must go beyond
  - establishing the local reality of our work

  And also begin
  - placing our work in the right time zone
- Two examples of linguistic “time zones”:
  - ISO 639-3 identifier [eng] is a zone covering all local varieties of English
  - GOLD identifier PastTense is a zone covering all local varieties of a past tense

A linguistic standard comes of age

- The perils of (cyber) train travel without standards:
  - Users miss the train altogether when the same thing has many different local names
  - Users experience a train wreck when many different things have the same local name
- The birth of linguistic information standards
  - “Each language is given a three-letter code on the order of international airport codes. This aids in equating languages across national boundaries, where the same language may be called by different names, and in distinguishing different languages called by the same name.” (Grimes 1974)
- The result: *ISO 639-3, Alpha-3 code for comprehensive coverage of languages* (published 2007-02-05)

A vision for interoperation in linguistics

- Part 1 — Global search
  - Interoperating over passive content
  - In a world of unlimited choice …
    - How can we ever find what we’re looking for?
- Part 2 — Global collaboration
  - Interoperating among active agents
  - When we realize the treasury is not very full …
    - How can we match supply with demand to fill it?

Part 1

Toward global search:

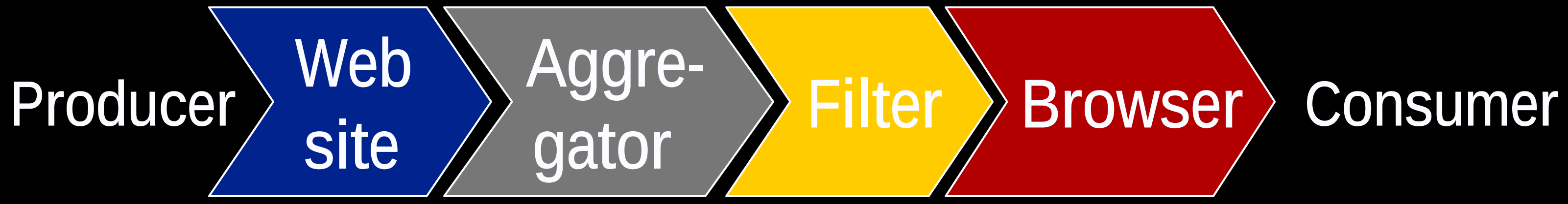
Interoperation over passive content

What linguists want

- Estimated size of today’s public Internet:
  - 100 million distinct web sites
  - 30 billion unique web pages
- “The language resource I’m looking for is probably out there somewhere, but …”
  - Which site should I look on?
  - How do I find the exact resource once I arrive?
  - It’s like searching for a needle in a haystack!
- The general strategy
  - Aggregate and Filter

Aggregate and Filter

- The value chain of Internet publishing:



Producer → Web site → Aggregator → Filter → Browser → Consumer

- Aggregators solve the “Which site?” problem by creating a single place to look.
- Filters address the “How do I find it?” problem by showing only the resources that match your criteria.

Two kinds of interoperation

- Shallow interoperation
  - Generic to all problem domains
  - Aggregates everything reachable via the ubiquitous HTTP infrastructure
  - Filters on the surface content of plain text
- Deep interoperation
  - Built for a specific problem domain
  - Uses domain-specific protocol to aggregate only what is relevant to the community
  - Uses domain-specific markup and vocabularies to filter on underlying concepts and structures

Good news and bad news

- Regarding shallow interoperation
  - It exists on a global scale (e.g. Google, Yahoo!) and is easy to support and use
  - But it gives poor results for language resources
    - Lots of noise (= low precision): words used in queries have many irrelevant senses
    - Lots of drop out (= low recall): words used in queries have synonyms and translations
- Regarding deep interoperation
  - It gives both high recall and high precision
  - But it takes more work to follow standards

Anchored by an aggregator

1. The community needs an aggregator for language resources to anchor its cyberinfrastructure.

- Provides a single authoritative inventory of every resource in the treasury of linguistic knowledge
  - Amasses the community’s assets into a single fund
  - The greater the fund, the greater the possible return on investment for the community
- An open API allows members of the community to invest the resources and earn interest by building services that add value for the community

Metadata standards

2. The community needs a metadata standard that describes the aggregated resources in a way that will support its filtering needs.

- The digital library community has developed a generic descriptive metadata standard:
  - Dublin Core Metadata Initiative
- The language resources community can simply augment the generic standard
  - Open Language Archives Community (OLAC) has done this: www.language-archives.org

OLAC metadata standard

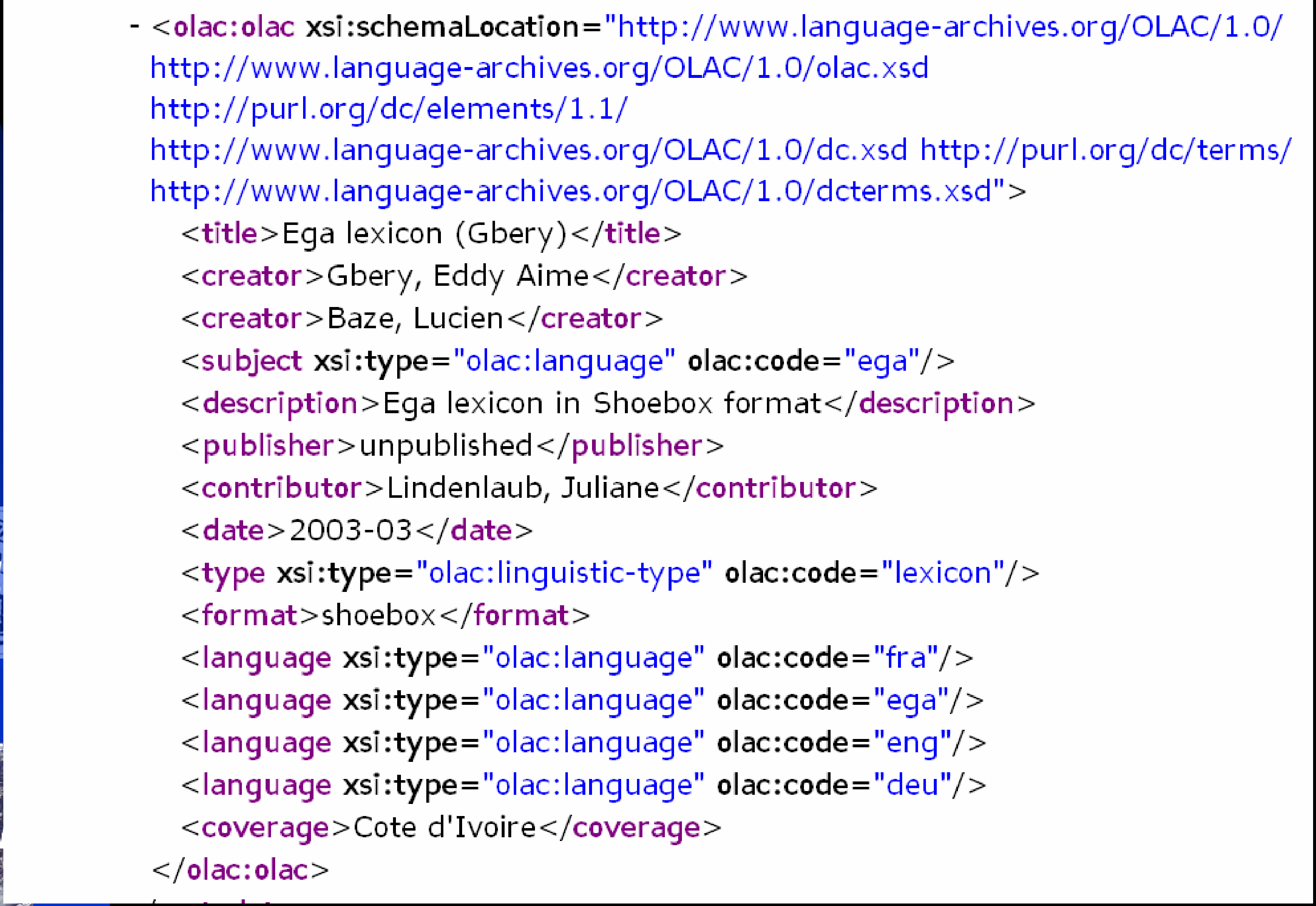
- Dublin Core metadata standard has:
  - Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title, Type
- OLAC adds extensions (with controlled vocabularies) specific to our community:
  - Language Identification (ISO 639-3), Linguistic Data Type, Linguistic Field, Participant Role, Discourse Type

Submission protocol

3. Institutions with language resources to share need an open protocol for submitting metadata to the aggregator.

- The digital library community has developed one:
  - Open Archives Initiative (OAI) protocol
- Our community can simply adapt it
  - OLAC has already done this
  - Specifies an exact syntax for resource description
  - 34 institutions are currently participating

A metadata record as submitted



```
- <olac:olac xsi:schemaLocation="http://www.language-archives.org/OLAC/1.0/
  http://www.language-archives.org/OLAC/1.0/olac.xsd
  http://purl.org/dc/elements/1.1/
  http://www.language-archives.org/OLAC/1.0/dc.xsd http://purl.org/dc/terms/
  http://www.language-archives.org/OLAC/1.0/dcterms.xsd">
    <title>Ega lexicon (Gbery)</title>
    <creator>Gbery, Eddy Aime</creator>
    <creator>Baze, Lucien</creator>
    <subject xsi:type="olac:language" olac:code="ega"/>
    <description>Ega lexicon in Shoebox format</description>
    <publisher>unpublished</publisher>
    <contributor>Lindenlaub, Juliane</contributor>
    <date>2003-03</date>
    <type xsi:type="olac:linguistic-type" olac:code="lexicon"/>
    <format>shoebox</format>
    <language xsi:type="olac:language" olac:code="fra"/>
    <language xsi:type="olac:language" olac:code="ega"/>
    <language xsi:type="olac:language" olac:code="eng"/>
    <language xsi:type="olac:language" olac:code="deu"/>
    <coverage>Cote d'Ivoire</coverage>
  </olac:olac>
```

Participating Archives

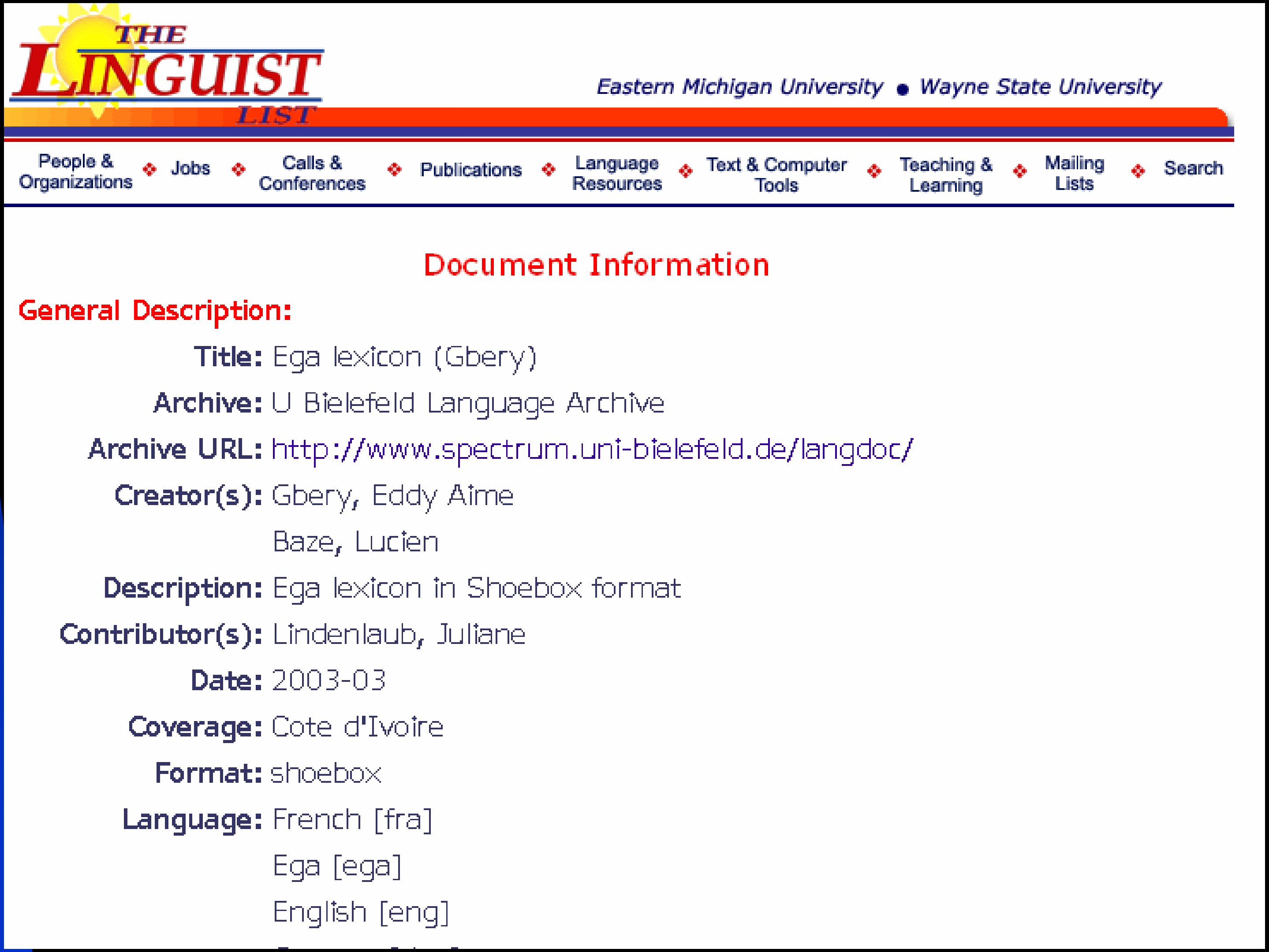
- Aboriginal Studies Electronic Data Archive
- Academia Sinica
- Alaska Native Language Center
- Archive of Indigenous Languages of Latin America
- ATILF Resources
- Berkeley Language Center
- Centre de Ressources pour la Description de l'Oral
- CHILDES Data Repository
- Comparative Corpus of Spoken Portuguese
- Cornell Language Acquisition Laboratory
- Dictionnaire Universel Boiste 1812
- DOBES catalogue (MPI, Nijmegen)
- Ethnologue: Languages of the World
- European Language Resources Association
- Laboratoire Parole et Langage
- Linguistic Data Consortium Corpus Catalog
- LINGUIST List Language Resources
- Natural Language Software Registry
- Online Database of Interlinear Text (ODIN)
- Oxford Text Archive
- PARADISEC
- Perseus Digital Library
- Research Papers in Computational Linguistics
- Rosetta Project 1000 Language Archive
- SIL Language and Culture Archives
- Surrey Morphology Group Databases
- Survey for California and Other Indian Languages
- TalkBank
- Tibetan and Himalayan Digital Library
- TRACTOR
- Typological Database Project
- University of Bielefeld Language Archive
- University of Queensland Flint Archive
- Virtual Kayardild Archive (Melbourne)

Harvesting protocol

4. Institutions who want to provide filtering services need an open protocol for harvesting metadata from the aggregator.

- The digital library community has developed one:
  - OAI Protocol for Metadata Harvesting
- Our community can simply adopt this
  - OLAC has already done this
  - Two institutions now provide search over the 30,000 aggregated resources
    - Linguist List and Linguistic Data Consortium

A metadata record as displayed



Document Information

General Description:

Title: Ega lexicon (Gbery)

Archive: U Bielefeld Language Archive

Archive URL: http://www.spectrum.uni-bielefeld.de/langdoc/

Creator(s): Gbery, Eddy Aime
Baze, Lucien

Description: Ega lexicon in Shoebox format

Contributor(s): Lindenlaub, Juliane

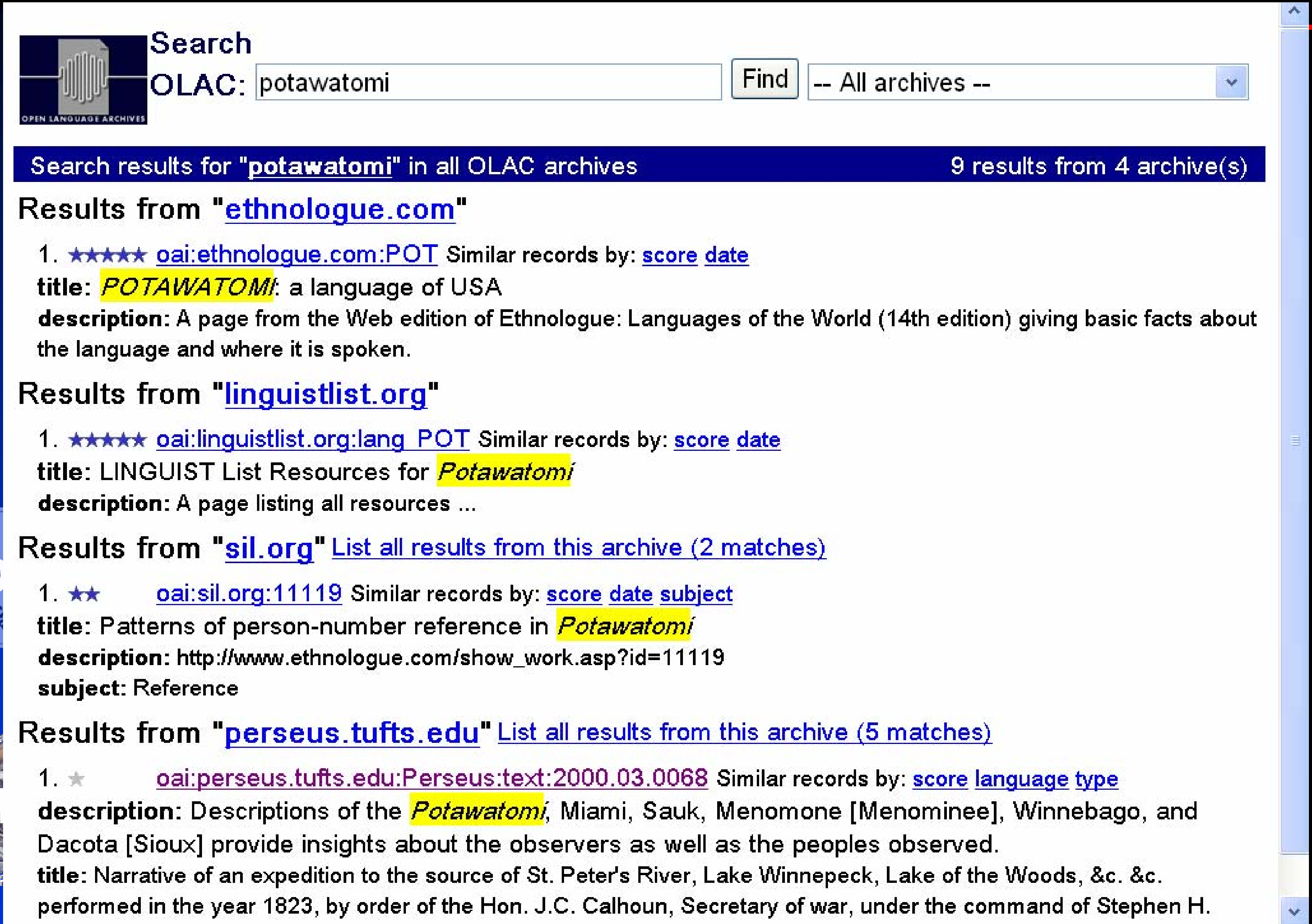
Date: 2003-03

Coverage: Cote d'Ivoire

Format: shoebox

Language: French [fra]
Ega [ega]
English [eng]

A listing of search results



Search OLAC: potawatomi [Find] -- All archives --

Search results for "**potawatomi**" in all OLAC archives 9 results from 4 archive(s)

Results from "ethnologue.com"

1. ★★★★★ oai:ethnologue.com:POT Similar records by: score date

title: *POTAWATOMI*: a language of USA

description: A page from the Web edition of Ethnologue: Languages of the World (14th edition) giving basic facts about the language and where it is spoken.

Results from "linguistlist.org"

1. ★★★★★ oai:linguistlist.org:lang_POT Similar records by: score date

title: LINGUIST List Resources for *Potawatomi*

description: A page listing all resources ...

Results from "sil.org" List all results from this archive (2 matches)

1. ★★ oai:sil.org:11119 Similar records by: score date subject

title: Patterns of person-number reference in *Potawatomi*

description: http://www.ethnologue.com/show_work.asp?id=11119

subject: Reference

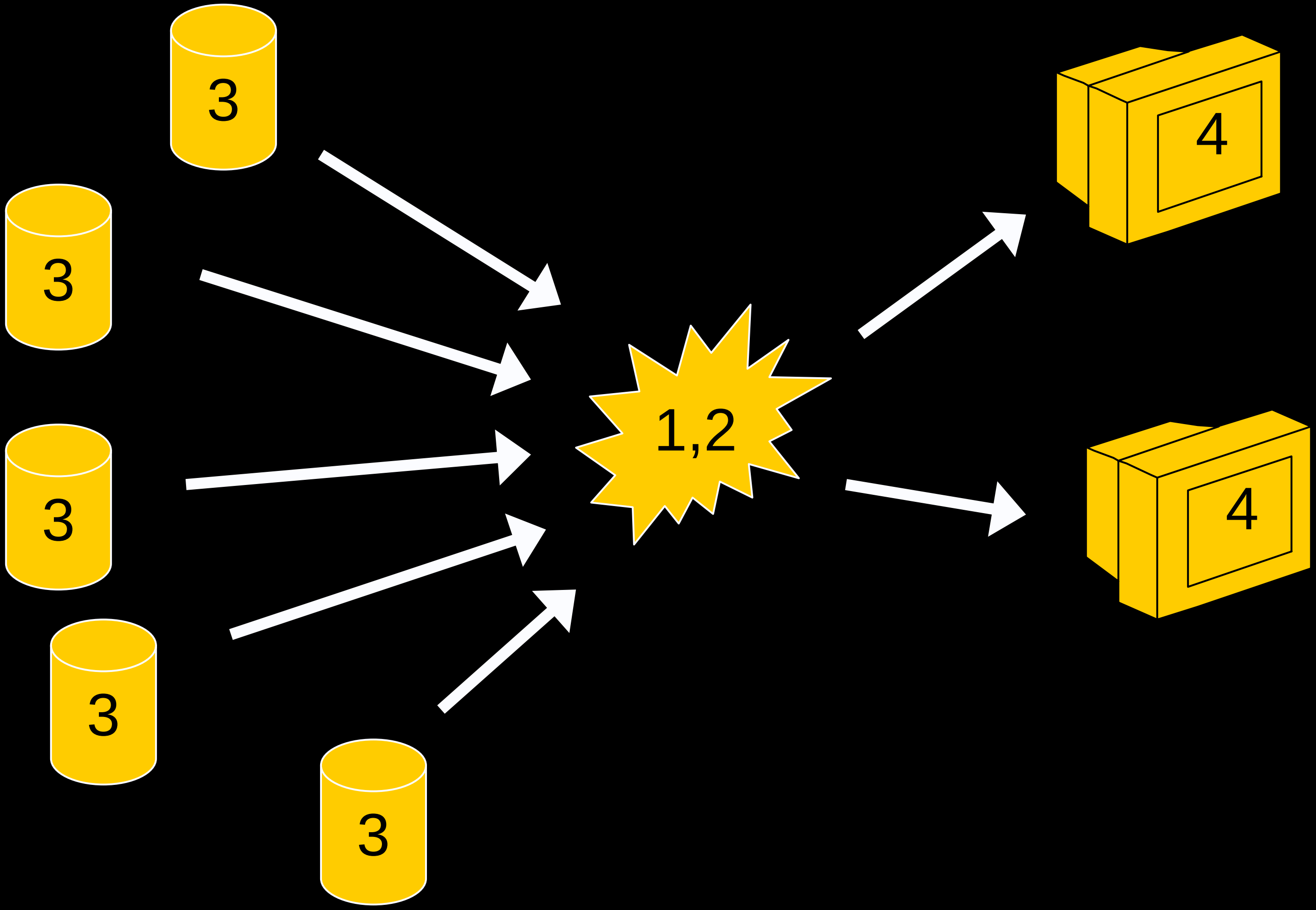
Results from "perseus.tufts.edu" List all results from this archive (5 matches)

1. ★ oai:perseus.tufts.edu:Perseus:text:2000.03.0068 Similar records by: score language type

description: Descriptions of the *Potawatomi*, Miami, Sauk, Menomone [Menominee], Winnebago, and Dacota [Sioux] provide insights about the observers as well as the peoples observed.

title: Narrative of an expedition to the source of St. Peter's River, Lake Winnepeck, Lake of the Woods, &c. &c. performed in the year 1823, by order of the Hon. J.C. Calhoun, Secretary of war, under the command of Stephen H.

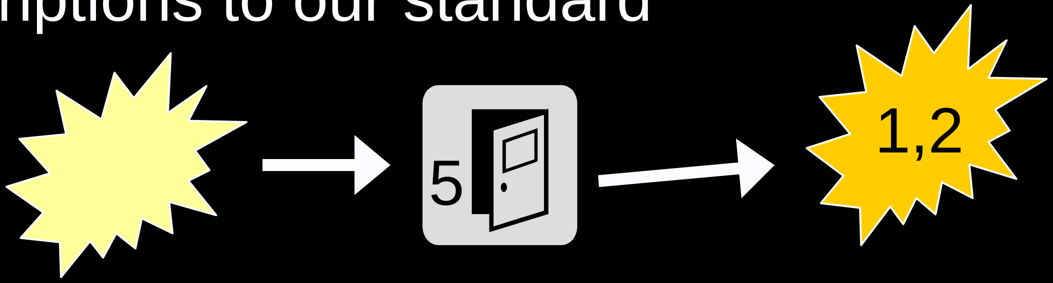
Existing OLAC infrastructure



Gateways

5. The cyberinfrastructure needs to include gateways to language resources catalogued in other academic aggregators.

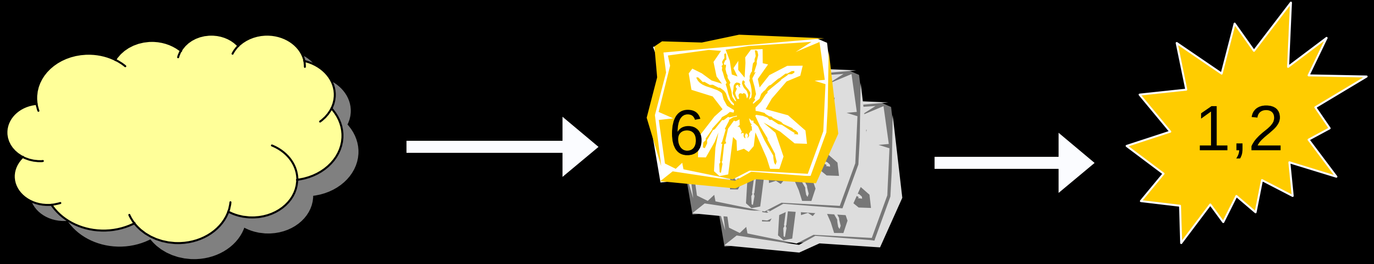
- General catalogs contain language resources
  - E.g., research libraries, institutional repositories
  - Our infrastructure needs gateways into these that discover language resources and “crosswalk” their descriptions to our standard



Spiders

6. The cyberinfrastructure needs to include spiders that discover language resources on the Web.

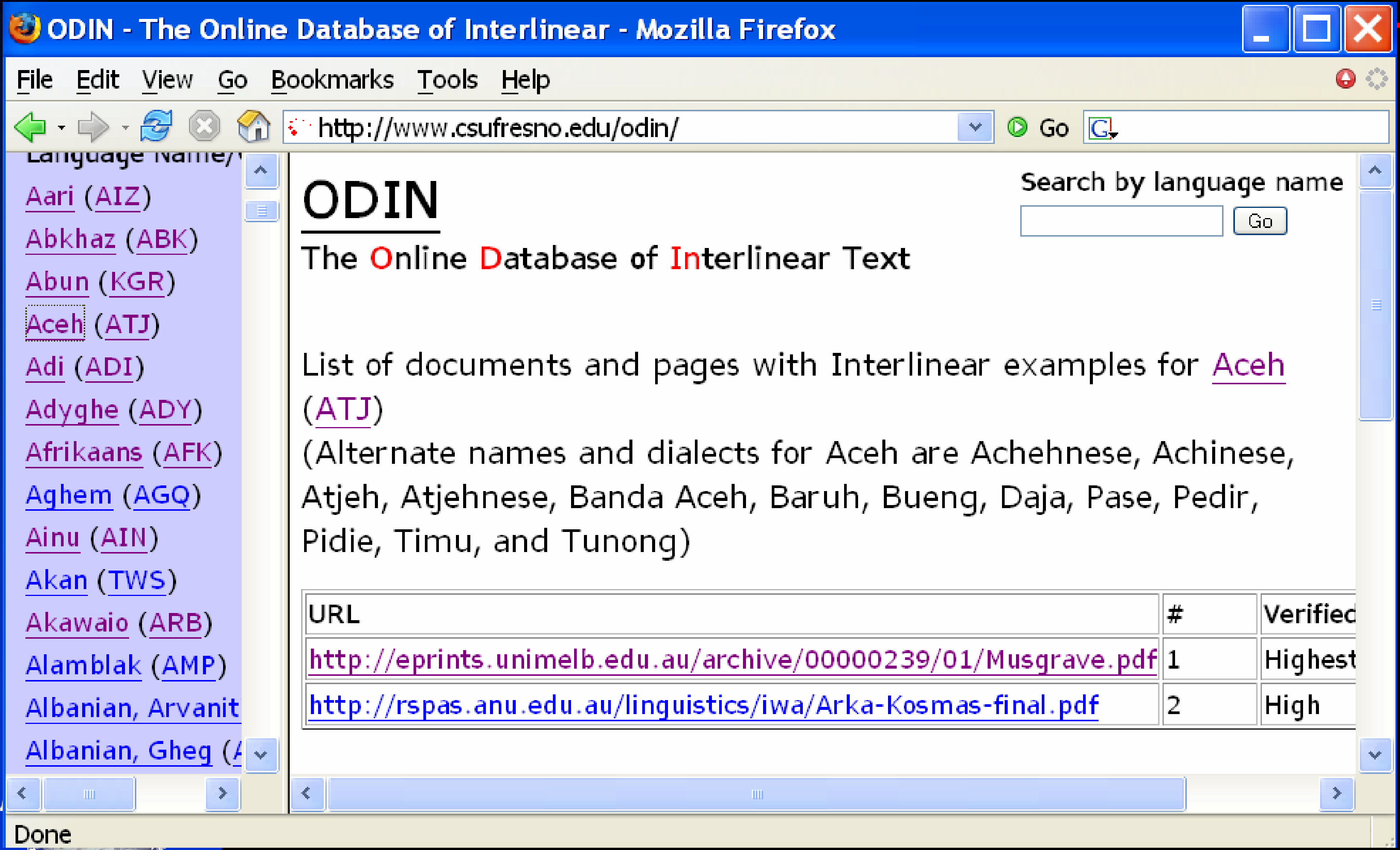
- Uncatalogued language resources abound on web
  - E.g. academic papers, language community blogs and web sites, pages in minority languages
- Our infrastructure needs spiders that will crawl the web to find and report them to our aggregator



The first linguistic spider exists

- ODIN: Online Database of Interlinear Text
  - Will Lewis, proceedings of 2003 E-MELD Workshop
- Methodology
  - Seed Google search with abbreviations for glosses
  - Keep URL if pattern for text-gloss-translation matches
  - Find language name in preceding context
- Service currently reports more than:
  - 34,000 instances of Interlinear Glossed Text examples
  - from over 700 different languages
  - in more than 2,200 different linguistic documents

What the user sees



ODIN - The Online Database of Interlinear - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.csufresno.edu/odin/
Go
Language Name/
Done

Aari (AIZ)
Abkhaz (ABK)
Abun (KGR)
Aceh (ATJ)
Adi (ADI)
Adyghe (ADY)
Afrikaans (AFK)
Aghem (AGQ)
Ainu (AIN)
Akan (TWS)
Akawaio (ARB)
Alamblak (AMP)
Albanian, Arvanit
Albanian, Gheg (A

ODIN

The Online Database of Interlinear Text

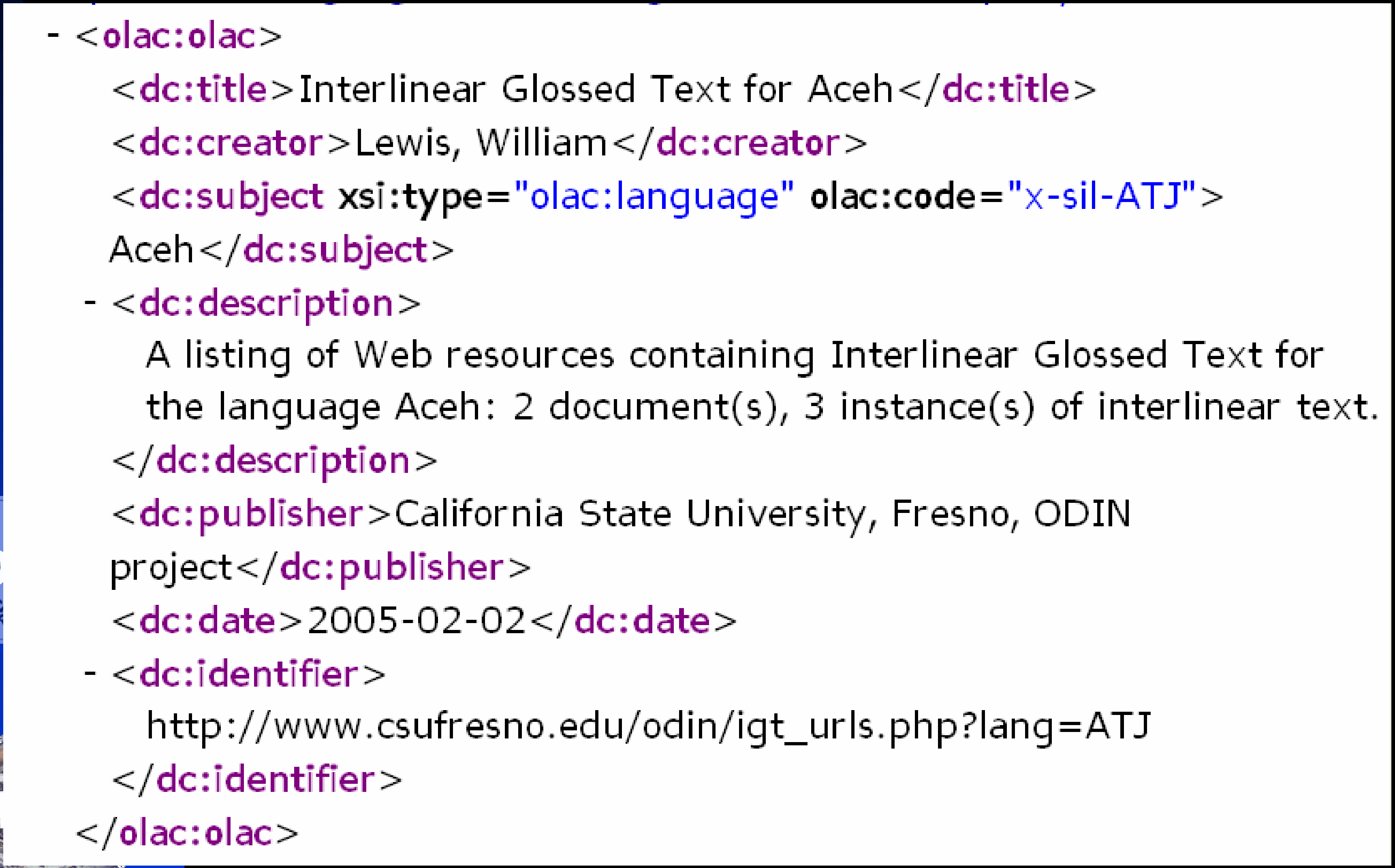
Search by language name

Go

List of documents and pages with Interlinear examples for Aceh (ATJ)
(Alternate names and dialects for Aceh are Achehnese, Achinese, Atjeh, Atjehnese, Banda Aceh, Baruh, Bueng, Daja, Pase, Pedir, Pidie, Timu, and Tunong)

URL	#	Verified
http://eprints.unimelb.edu.au/archive/00000239/01/Musgrave.pdf	1	Highest
http://rspas.anu.edu.au/linguistics/iwa/Arka-Kosmas-final.pdf	2	High

What the aggregator sees



```
- <olac:olac>
    <dc:title>Interlinear Glossed Text for Aceh</dc:title>
    <dc:creator>Lewis, William</dc:creator>
    <dc:subject xsi:type="olac:language" olac:code="x-sil-ATJ">
    Aceh</dc:subject>
  - <dc:description>
      A listing of Web resources containing Interlinear Glossed Text for
      the language Aceh: 2 document(s), 3 instance(s) of interlinear text.
    </dc:description>
    <dc:publisher>California State University, Fresno, ODIN
    project</dc:publisher>
    <dc:date>2005-02-02</dc:date>
  - <dc:identifier>
      http://www.csufresno.edu/odin/igt_urls.php?lang=ATJ
    </dc:identifier>
  </olac:olac>
```

Subcommunities

7. Subcommunities need to establish conventions within the metadata standard in order to support specialized filtering, gateways, and spiders.

- Subcommunities have more specialized filtering needs than the OLAC standard supports
  - Just as OLAC was immediately successful because it specialized existing standards (OAI, DC)
  - Subcommunities can do the same by specializing the OLAC standard
  - Then build services that exploit the specializations

Examples of subcommunities

- LL-MAP project using extensions to the Coverage element for geospatial coordinates
  - ▸ *E.g.*, <dc:coverage xsi:type="dcmi:point">name= Perth, W.A.; east=115.85717; north=-31.95301 </dc:coverage>
- Tomorrow's workshop: *Applying Cyber-infrastructure to the Language Sciences: A Case Study for Language Acquisition*
- Can be defined by a fixed metadata value
  - ▸ *E.g.*, <dc:format>LIFT 1.0</dc:format>

Datum-level search

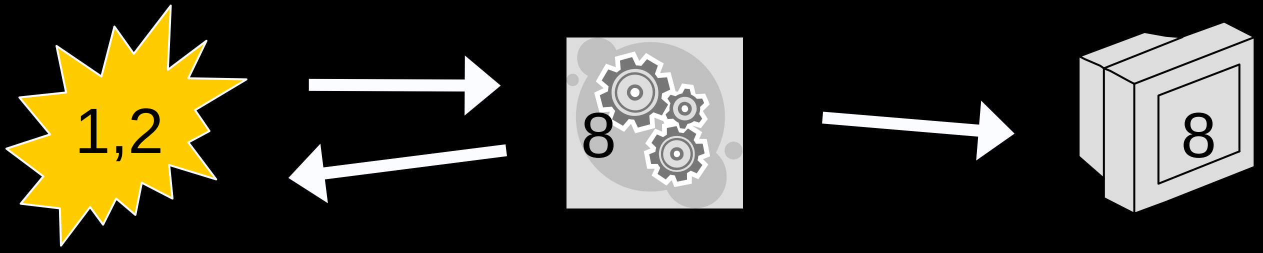
8. Subcommunities need to establish content standards in order to support datum-level search over relevant resources harvested from the aggregator.

- We also want to search across resources for matching data within the content
- Deep search implies that the subcommunity interested in a particular content type must establish standards for markup or vocabularies
- Papers for two past EMELDs illustrated this
  - 2003, Mapping lexicons to GOLD standard
  - 2004, Mapping interlinear glossed texts

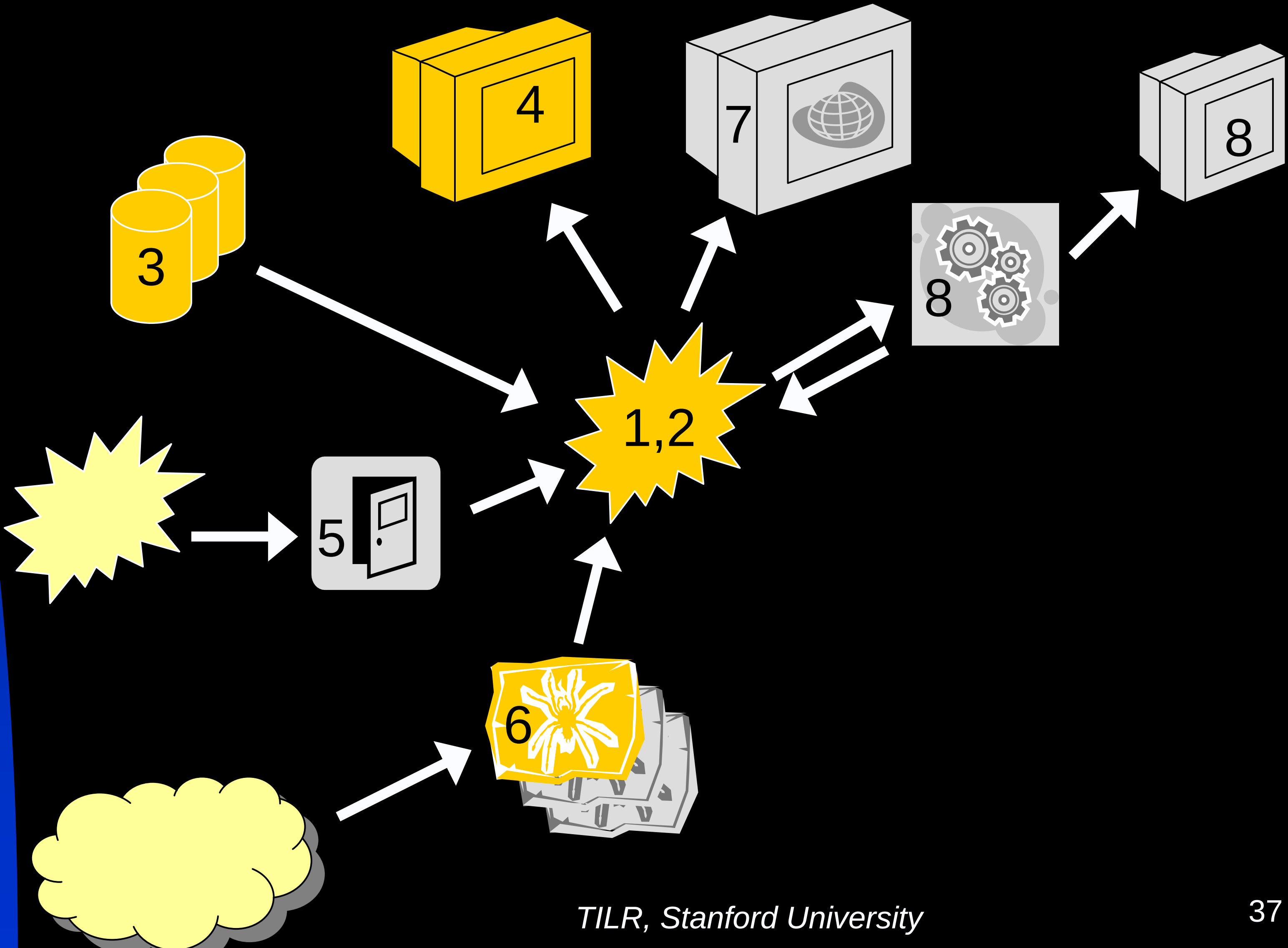
Implementing datum-level search

- Aggregate and filter at a more granular level:

1. Query aggregator for all subcommunity resources
2. Harvest the resources themselves
3. Parse content into a content-specific database
4. Implement search service over that database
5. Submit descriptions of the reports generated



Cyberinfrastructure for linguistics



DTSL

EMELD

Electronic Metastructure for Endangered Languages Data

Part 2

Toward global collaboration:

Interoperation among active agents

What linguists are going to want

- Today the big challenge is to pull together a cyberinfrastructure that amasses all known language resources
- Once we get it, we will realize how empty the global treasury still is.
  - “The information I’m looking for isn’t there!”
- The next big question will be
  - “How do we fill the huge gaps in the global treasury of linguistic knowledge?”

The world is flat

- Thomas Friedman, 2005
  - *The World is Flat: A brief history of the 21st century*
- 1492: Columbus
  - Sailed west to find “the countries of India”
  - He arrived demonstrating the world was round
  - Except he’d actually gone to America
- 2004: Friedman
  - Flew east to India and really got there
  - Except what he found there was a lot of America
  - The world isn’t round any more; it’s been flattened

The stages of globalization

- Globalization 1.0
  - 1492: Countries begin to globalize
- Globalization 2.0
  - 19th century: Industrial age makes it possible for companies and organizations to globalize
  - Global riches of knowledge in world class libraries
- Globalization 3.0
  - 21st century: Information age creates a flattened world in which individuals can globalize
  - Global riches of knowledge in every web browser

Three things converging

1. Playing field
   - Innovations in the last 15 years have produced a global, Web-enabled playing field for collaboration.
2. Processes
   - After a decade or so business processes finally changed to take advantage of new technologies and achieve the great productivity breakthrough.
3. People
   - Three billion people who had been frozen out of the field are suddenly able to plug and play with everybody else.

The triple convergence

- *"It is this triple convergence—of new players, on a new playing field, developing new habits and processes for horizontal collaboration—that I believe is the most important force shaping global economics and politics in the early 21st century. … The scale of the global community that is soon going to be able to participate in all sorts of discovery and innovation is something the world has simply never seen before." (pp. 181–182)*

Paradigm shift in progress

- Traditional approach to knowledge management
  - Experts collect existing knowledge and organize it into a knowledge base
- The emerging approach
  - Knowledge is a common good that is collaboratively produced.
  - Decreasingly produced individually but increasingly by distributed and virtually organized groups
- Paradigm shift
  - From knowledge warehouse management
  - To collaborative knowledge production

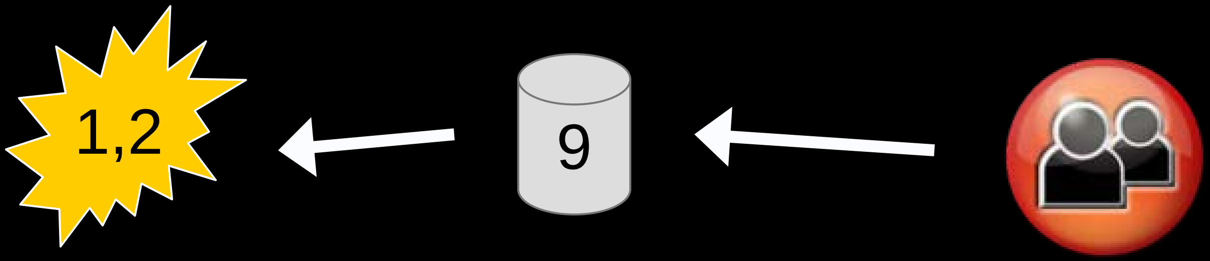
Open repositories

9. Individuals need an open method for submitting language resources into repositories that are plugged into the aggregator.

- Remove the archiving bottleneck by setting up self-service digital archives for language resources
- Institutional repositories are already mainstream
- *E.g.*, DSpace at 8 of top 30 universities in world:
  - University of Cambridge
  - MIT
  - Cornell University
  - Australian National Univ.
  - National Univ. of Singapore
  - University of Melbourne
  - University of Toronto
  - University of Michigan

Self-archiving of language resources

- Self-archiving workflow
  1. Contributor uses web forms to fill in metadata and upload content to the repository
  2. Curator verifies metadata and fitness for collection
  3. Resource is immediately available to the aggregator



- Prescription for our community
  - ▸ Dspace already supports the OAI protocol; just add support for OLAC metadata

Collaboration with amateurs

10. The cyberinfrastructure needs to include services that support generation and refinement of resources through collaboration with amateurs.

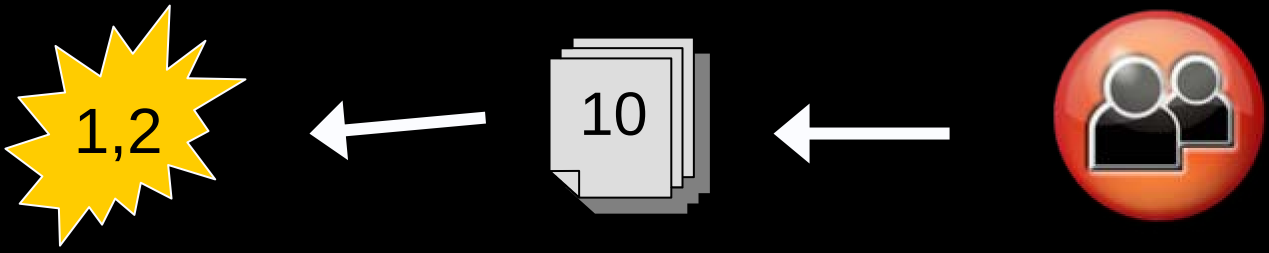
- Recent essay in *Nature* about the growing use of collaborative methods in science
  - "The Common Good," by Philip Ball
- For example,
  - Project Gutenberg
  - Wikipedia
  - NASA Mars Clickworkers

The next great Pro-Am movement?

- The universe is too big for professional astronomers to be looking everywhere all the time
  - They've increased coverage by an order of magnitude by embracing collaboration with amateurs
  - Chris Anderson, *The Long Tail,* Ch. 5, "The new producers: Never underestimate the power of a million amateurs with keys to the factory"
- The great challenge for 21st century linguistics
  - Everyone on earth has personal, linguistic knowledge that is of interest to professionals
  - Exploit the global collaborative playing field to enlist everyone in the quest for that knowledge

Possible applications

- Language documentation workflow: create, transcribe, translate, annotate, evaluate, refine
- Wiktionary
- Elicitation schedules
- Verifying output of automated services like spiders, taggers, parsers, language identifiers
- In all cases the results should be submitted to the aggregator so they will be found in searches:



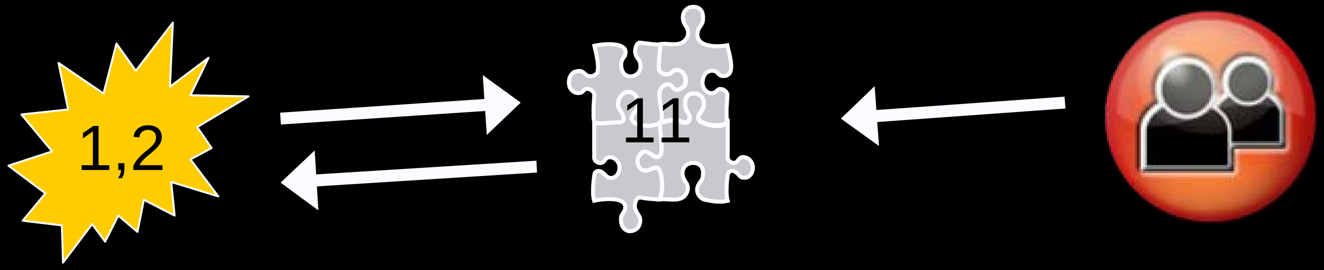
Supply and demand

11. The community needs an infrastructure for matching the demand for work to be done with the supply of people who could do it.

- Harnessing the collaborative work force is a matter of matching supply and demand
- The interactive web has excelled at
  - Matching consumers to niche products
    - Amazon, Rhapsody, Netflix
  - Matching individual buyers and sellers
    - eBay, eHarmony

Supply and demand for language resources

- We have a huge demand to:
  - Gather, convert, transcribe, translate, segment, annotate, analyze, describe, evaluate
  - We need an infrastructure for expressing this demand and fitting it to the supply of volunteers
- It should link to the aggregator in order to
  - Extract inferences about next workflow steps
  - Inject descriptions of work needing to be done



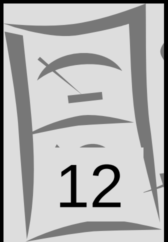
Authority and reputation

12. The community needs protocols for measuring the authority of resources and the reputation of contributors within the cyberinfrastructure.

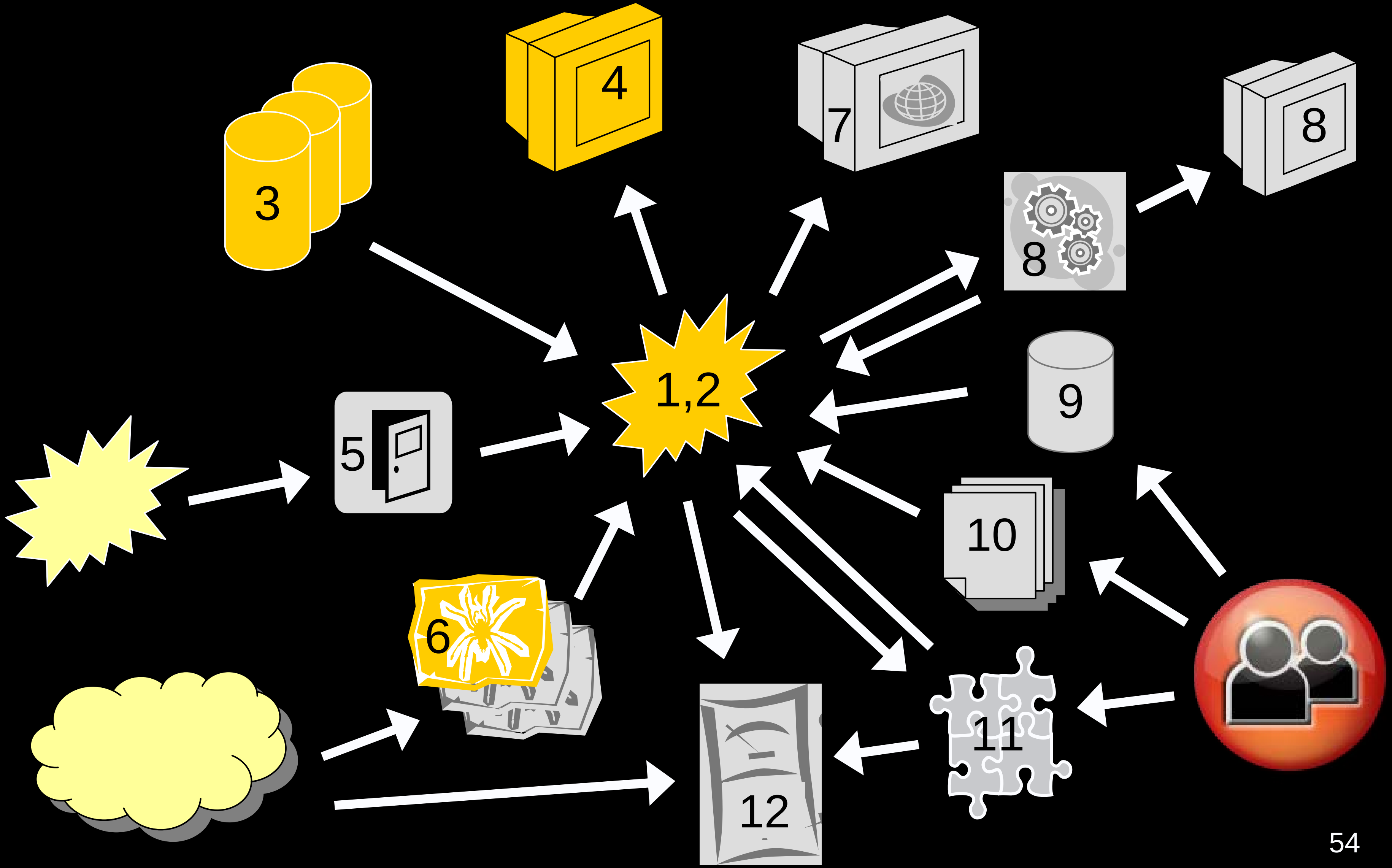
- In economy of scarcity, authority conferred by
  - Prestige of journal, publisher, author’s institution
  - Author’s degree and tenure status
- In economy of abundance, new metrics are emerging from the “collective intelligence”
  - For the authority of resources
  - For the reputation of contributors

Providing incentive and gratification

- Measuring the authority of resources
  - User ratings, user comments, % found this helpful, Google page rank, page views, outgoing links, incoming links, number of edits and editors
- Measuring the reputation of contributors
  - Buyer/seller ratings, number of contributions, authority of contributions
  - Reputational recognition (not financial) is what will drive the economy of linguistic abundance
- Infrastructure needs meters for the above
  - With inputs from the aggregator, supply-demand matcher, incoming links from web



Cyberinfrastructure for linguistics



3
4
7
8
8
1,2
5
9
10
6
11
12