Towards a comprehensive *Languoid* Catalog

Michael Cysouw, MPI-EVA (cysouw@eva.mpg.de)
Jeff Good, University at Buffalo (jcgood@buffalo.edu)

The objects of study

- Any properly structured catalog requires a *rigorous, generally applicable definition* of what kinds of objects should be contained within it
- The concept of "*language*" is clearly not amenable to such a definition
- Our main point here: we should aim for a comprehensive *languoid* catalog using *doculects* as the basis for defining them.

Background: model and view

- *Model*: representation of the data being cataloged
  - Models may need to be quite complex
- *View*: rendering the model to facilitate user interaction with the data
  - Views should be simple and intuitive
- Here, we are focusing on a *model* for the objects in a language catalog

What is a Languoid?

Languoids are *language-like entities*, including

- All kinds of lects:
  - language, dialect, sociolect, idiolect, stylistic register, etc.
- Genealogical groupings:
  - all levels, from dialect cluster to stock
- Geographic groupings:
  - sprachbund, spread zone, macro area, climate zone, continent, etc.

Why languoid?

- To allow us to move forward to catalogue "languages" while avoiding the insoluble problem of deciding what a "language" is
- A languoid can be catalogued separately from the specification as to what kind of languoid it is
  - Dialect or language?
  - Language or small family?
  - Genealogical or areal group?
  - Different register or different language?

Defining Languoids

- Languoids are defined as a set of languoids
- Recursion ends at doculects
- *Doculect*: variety as instantiated by any available documentation
  - Grammatical description (grammar, article)
  - Dictionary, wordlist
  - Inscription, transcription, recording
  - Description of personal knowledge
  - Language notes in a traveller's diary
  - Name given in an ancient text, or in a census

Catalogue Components

- Basic structure of a languoid
  - Unique identifier
  - Specification of the authority claiming the existence of the languoid
  - Name as used by the authority
  - Specification of what other languoids it encompasses (ideally down to the doculect)
  - …

Contested data

- Under this conception, whether or not a languoid exists will not be controversial
- If it's mentioned in a citable source, it's a languoid—and we should catalog it
- Controversies will arise on issues like:
  - which languoids make sense?
  - how do similar languoids relate to each other?
  - what kind of languoid is it? (language, dialect, stylistic variant, etc.)
  - how do names relate to languoids?

Languoid names: Glossonyms

- Basic structure of a glossonym
  - Unique identifier
  - A text string
  - Language the string is written in (e.g., English, German, Spanish)
  - Authority using this name
  - ...

Glossonym synonymy

- Glossonyms can then be grouped into *glosso-synonym sets* of distinct glossonyms referring to the same languoid
  - e.g. “German”, “Deutsch”, “Немецкий”
- One glosso-synonym set may be associated with distinct languoids
  - e.g. “Altaic” and its glosso-synonyms

Glossonym homophony

- Distinct glossonyms associated with the same text string in the same language get different IDs
- So, the Nilo-Saharan language glossonym "Aka" would be represented distinctly from the Bantu language glossonym "Aka"
- More tricky is a situation like "Maku", which is both a lect ("Yuhup") and a small family including the lect ("Puinavean", including Yuhup)

Implementation

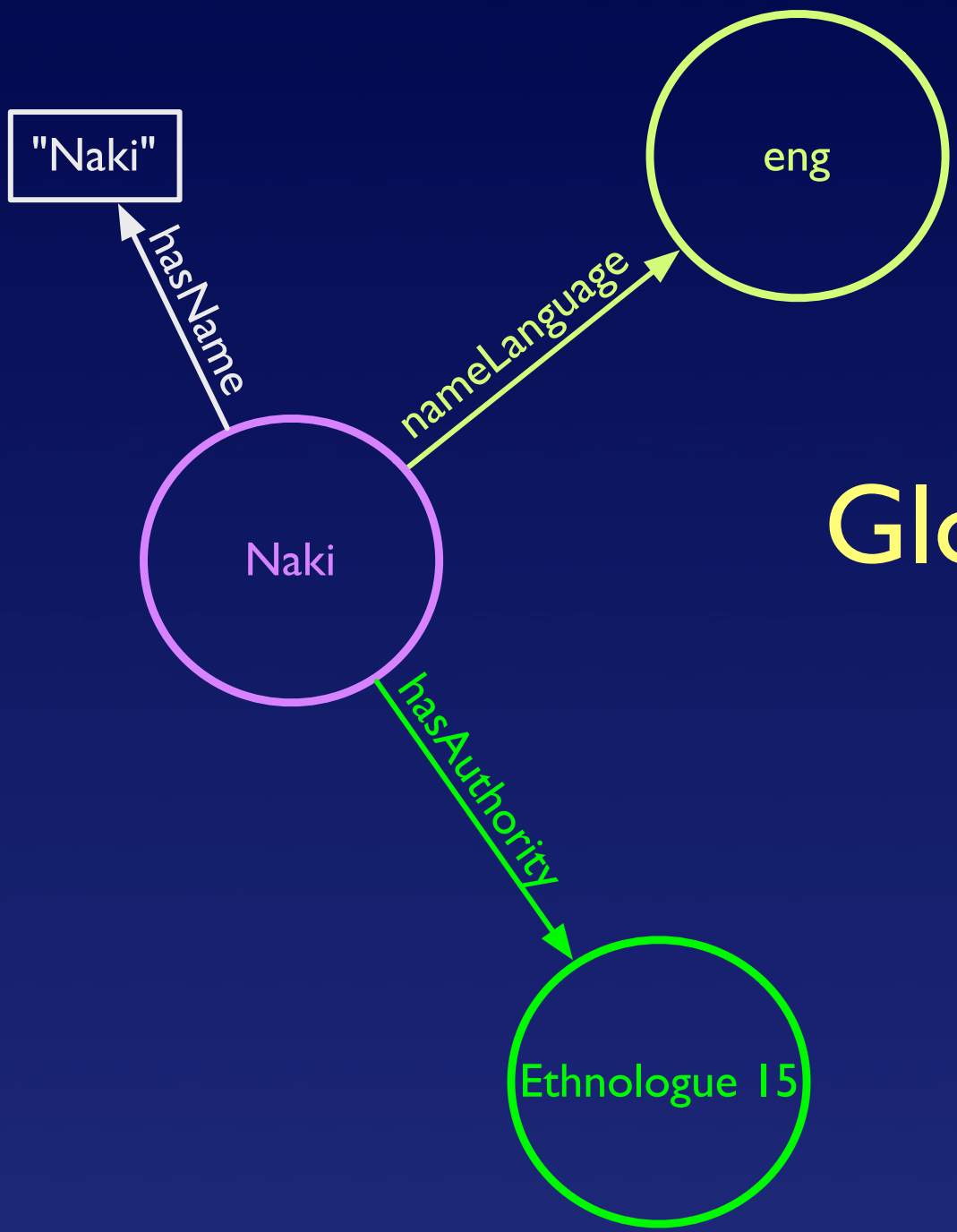
- We believe we should assume at the outset that
  - We do not know all the kinds of information we may want to associated with languoids, glossonyms, and authorities
  - We do not know all the ways we might want to link languoids, glossonyms, and authorities and all other information to each other

Implementation

- Therefore, we should seek implementations which
  - Give us flexibility to add new kinds of information to the database without "breaking" it
  - Allow for open-ended ways of referring to and grouping the different entities in the database

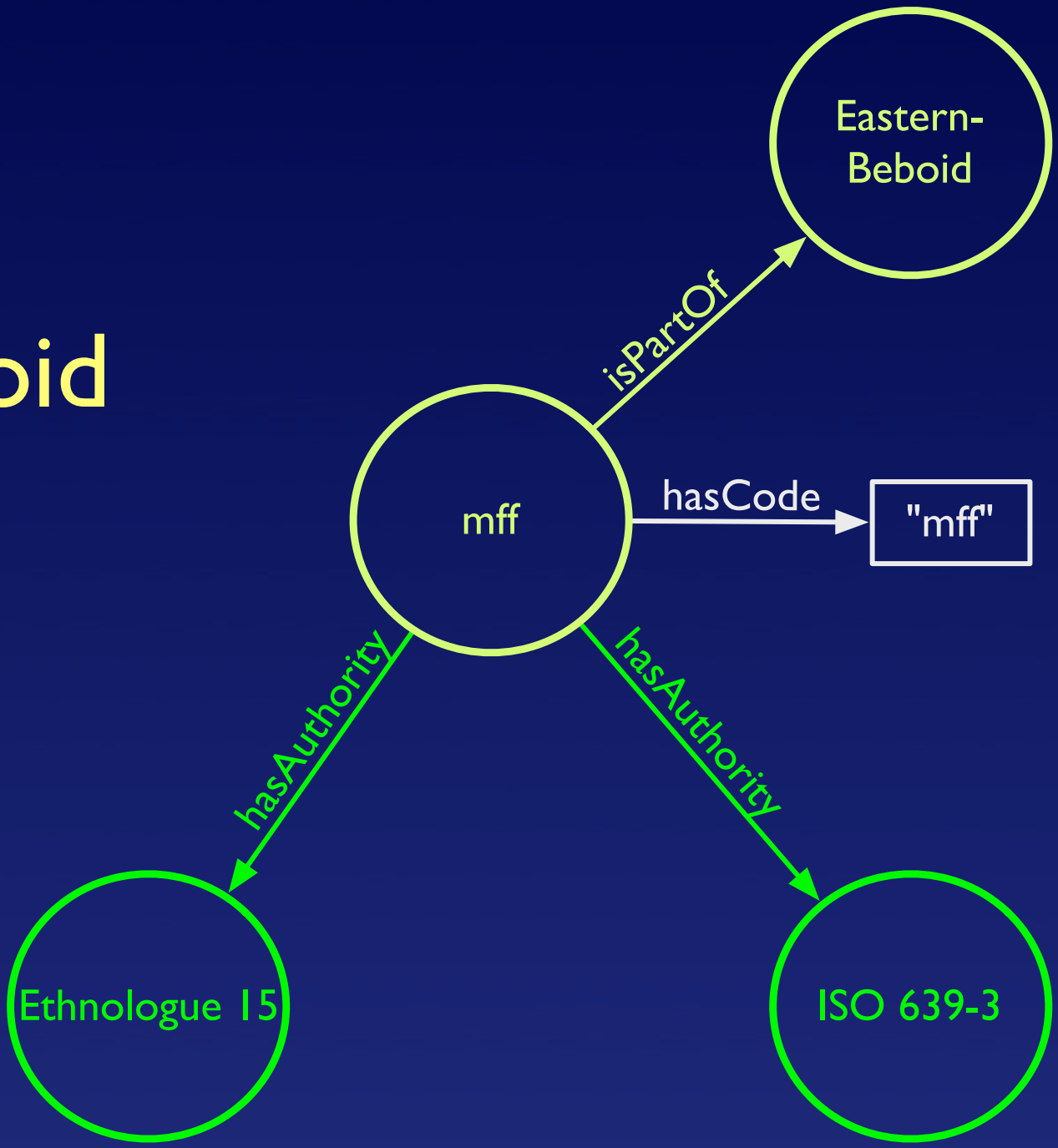
Graphs

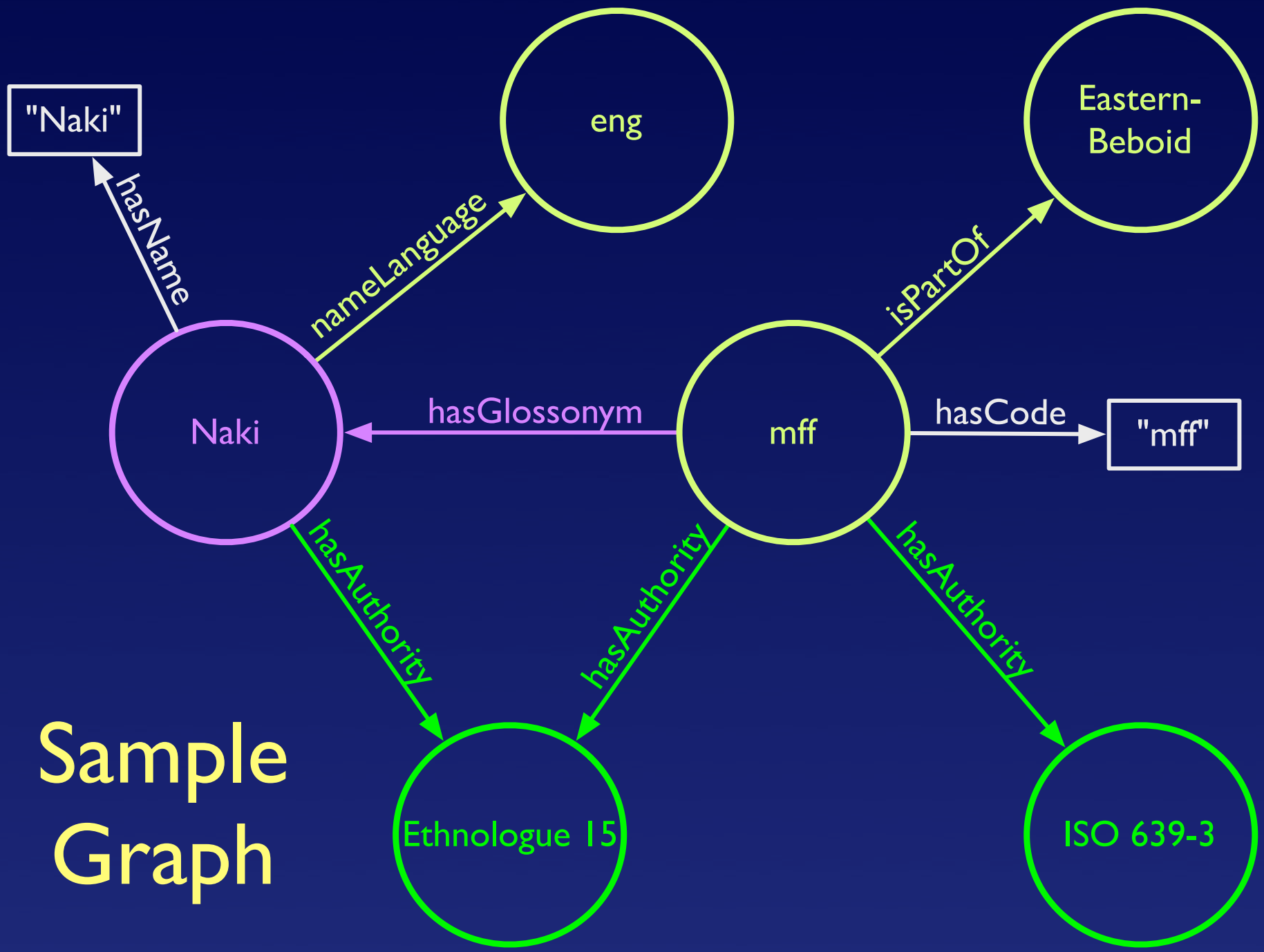
- We believe a graph-based representation is better suited for a project like this than a table-based one
- The *Resource Description Framework (RDF)* method of encoding graphs is one prominent technology that is up to this task


"Naki"
hasName
eng
nameLanguage
Naki
hasAuthority
Ethnologue 15

Glossonym

Languoid


Eastern-
Beboid
isPartOf
mff
hasCode
"mff"
hasAuthority
hasAuthority
Ethnologue 15
ISO 639-3



Sample Graph

Gaps

- The sort of system we are proposing would not require that all information be filled in for all languoids
- Rather, it would tolerate missing information, which would simply mean that the relevant part of the graph is missing
- Often, one might even want to propose a temporary "dummy" languoid (e.g. linking an authority to a glossonym)

Distributed information

- A graph-based system is also amenable to information being distributed across multiple sites
- Each site would simply store subgraphs about which it has information
- These could be joined, as needed, into a larger graph

Requirements

- The most important, and stringent, requirement is a unified system for the creation and registration of unique IDs
- Basically, we need four kinds of IDs
  - Languoids
  - Doculects
  - Glossonyms
  - Authorities