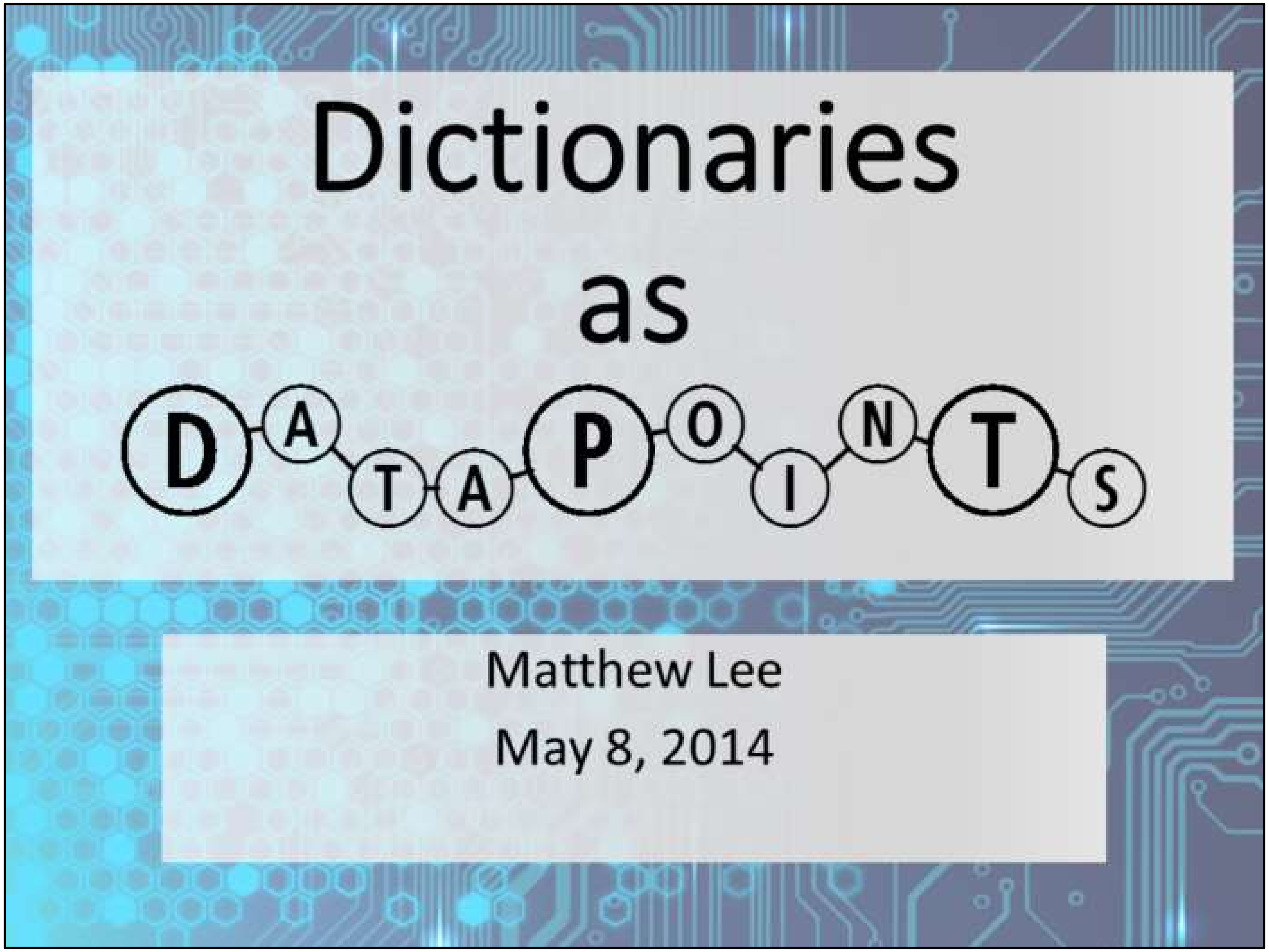


Dictionaries as DATAPOINTS

Matthew Lee

May 8, 2014

Matthew Lee: Who am I?

- Language Technology Specialist

  Member of SIL Cameroon

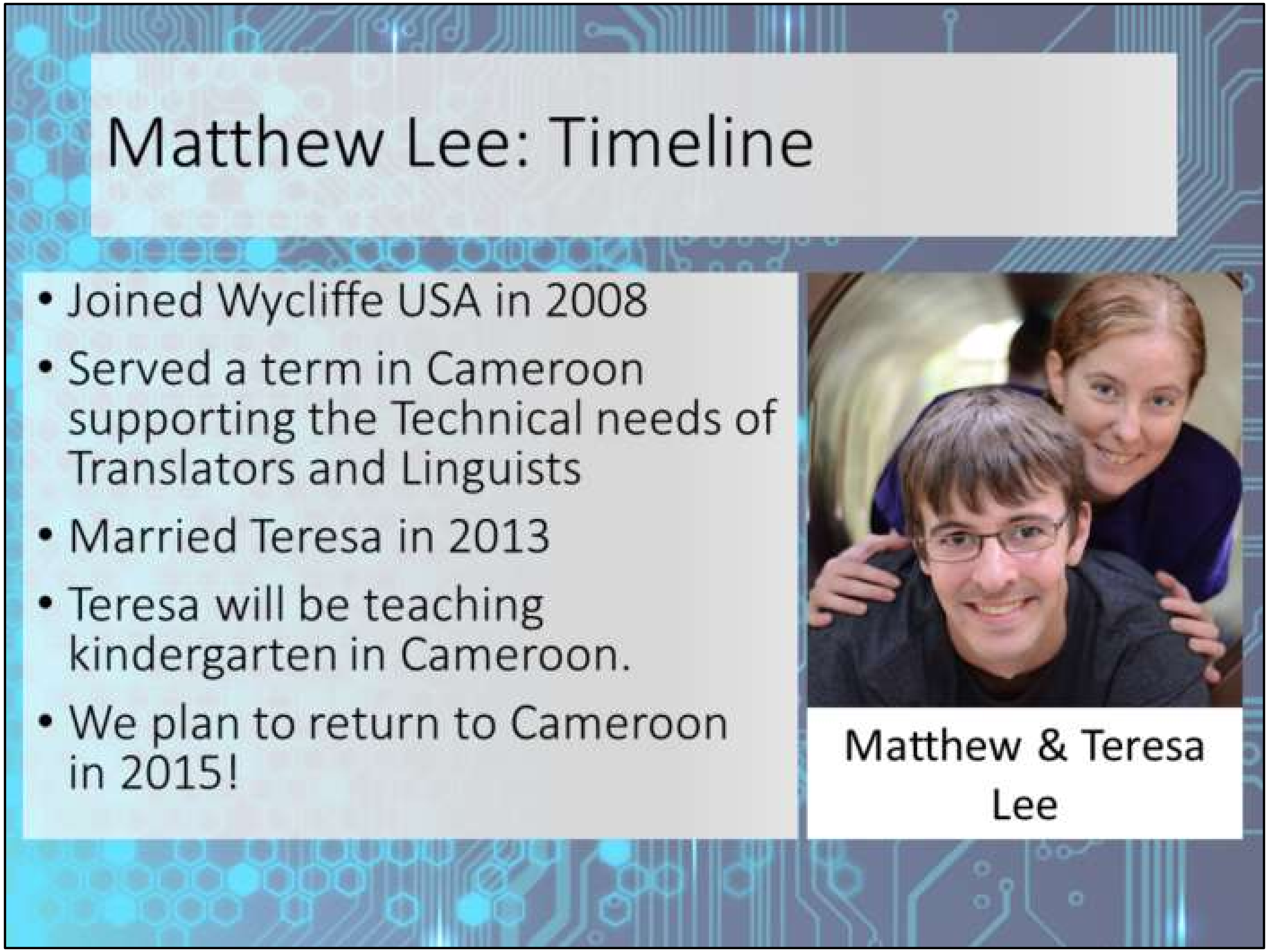
SIL CAMEROON

G·I·A·L

- Pursuing Master's Degree in Descriptive Linguistics

- BS: Integrated Science & Technology
- BA: Philosophy

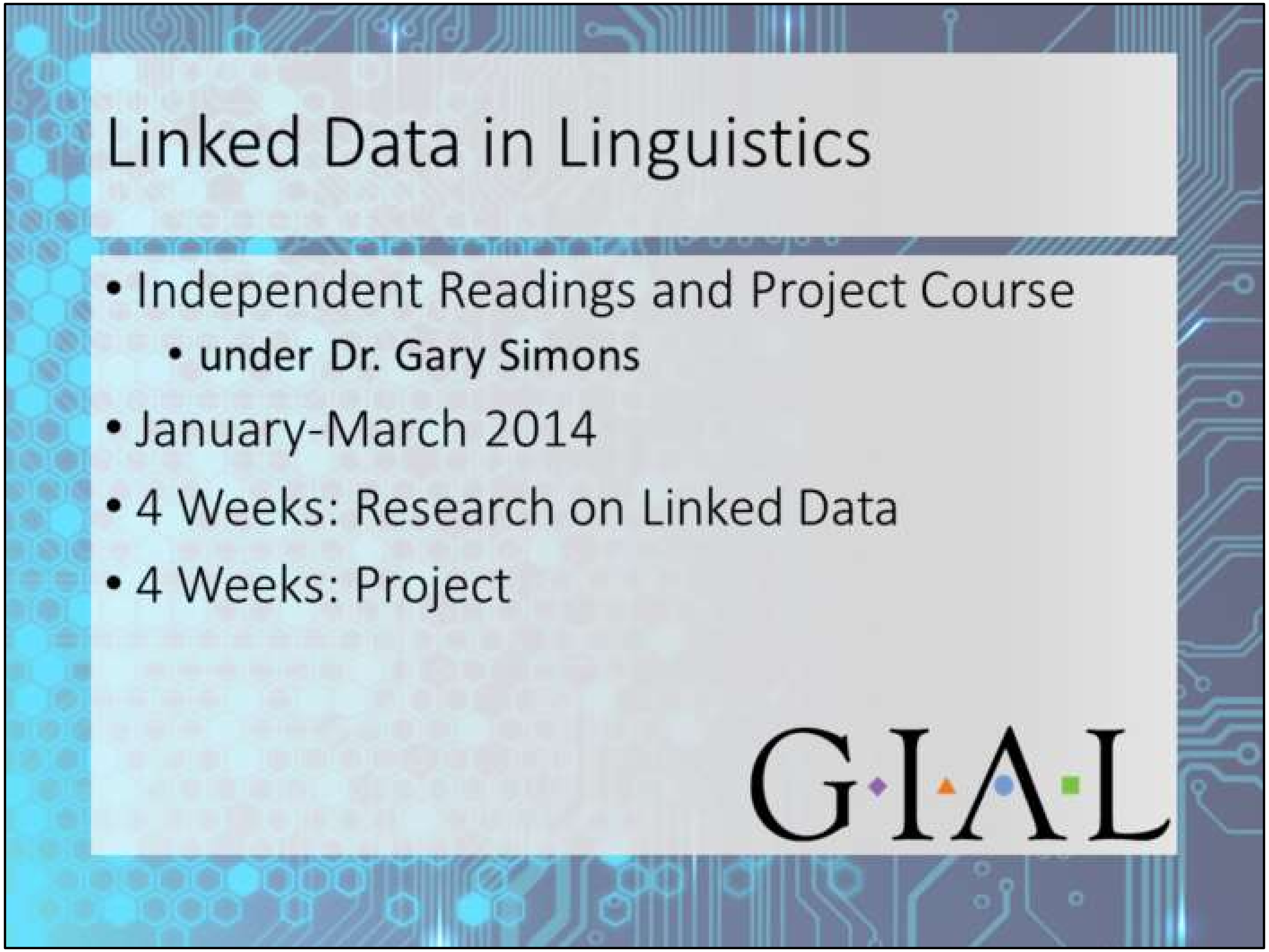
JAMES MADISON UNIVERSITY.



Matthew Lee: Timeline

- Joined Wycliffe USA in 2008
- Served a term in Cameroon supporting the Technical needs of Translators and Linguists
- Married Teresa in 2013
- Teresa will be teaching kindergarten in Cameroon.
- We plan to return to Cameroon in 2015!

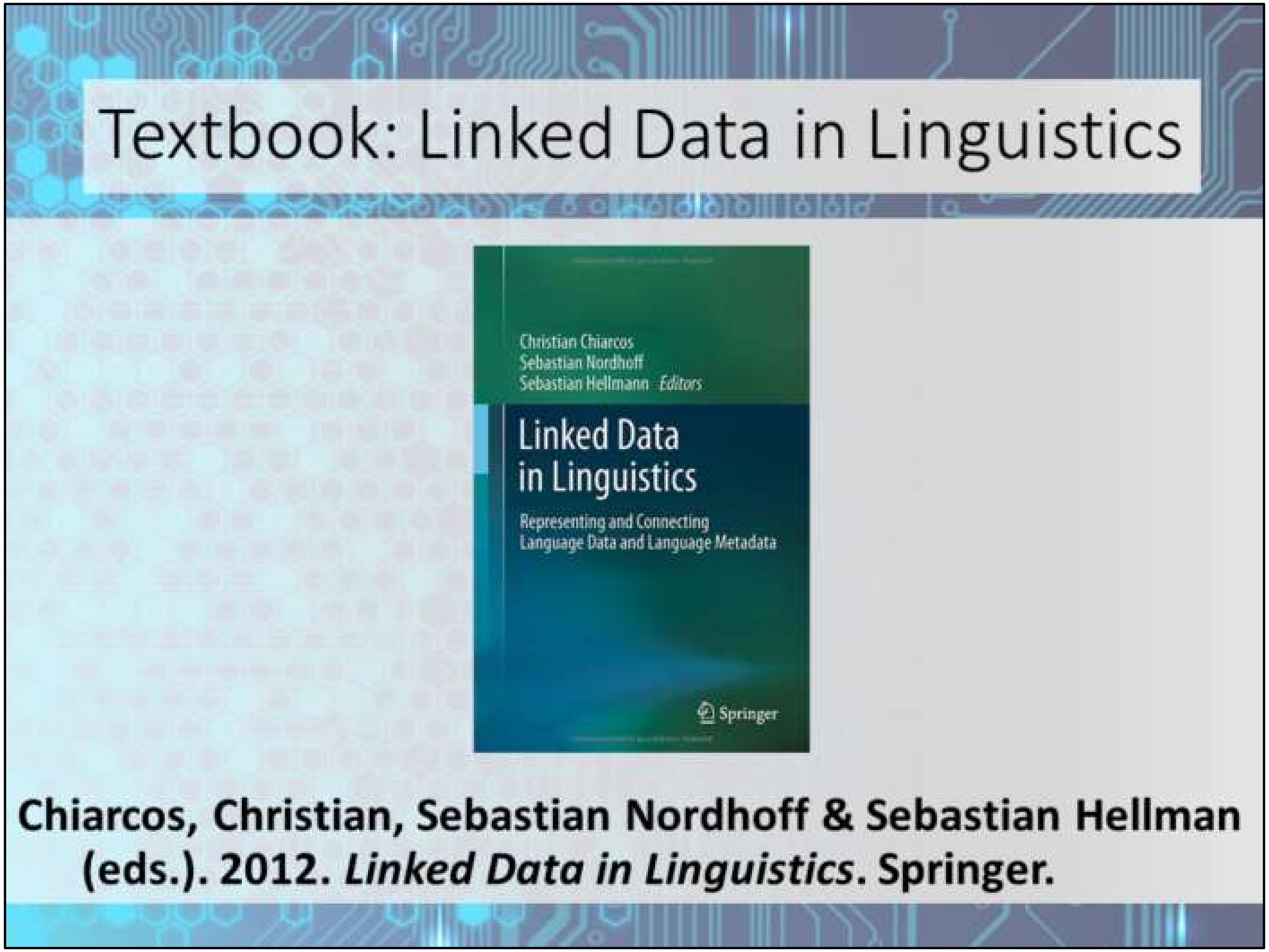
Matthew & Teresa Lee



Linked Data in Linguistics

- Independent Readings and Project Course
  - under Dr. Gary Simons
- January-March 2014
- 4 Weeks: Research on Linked Data
- 4 Weeks: Project

Independent Study

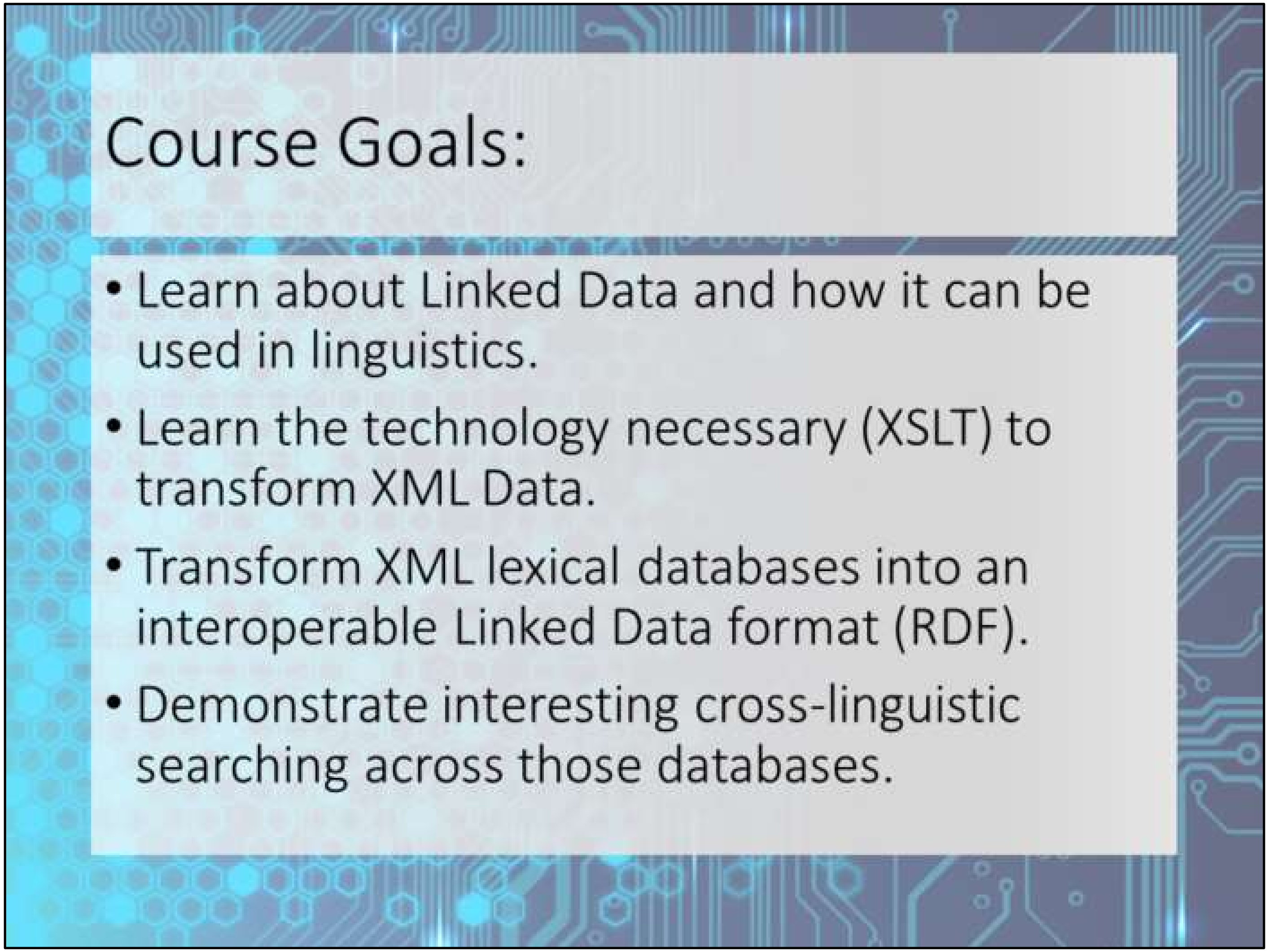


Textbook: Linked Data in Linguistics

Christian Chiarcos
Sebastian Nordhoff
Sebastian Hellmann Editors
Linked Data in Linguistics
Representing and Connecting Language Data and Language Metadata
Springer

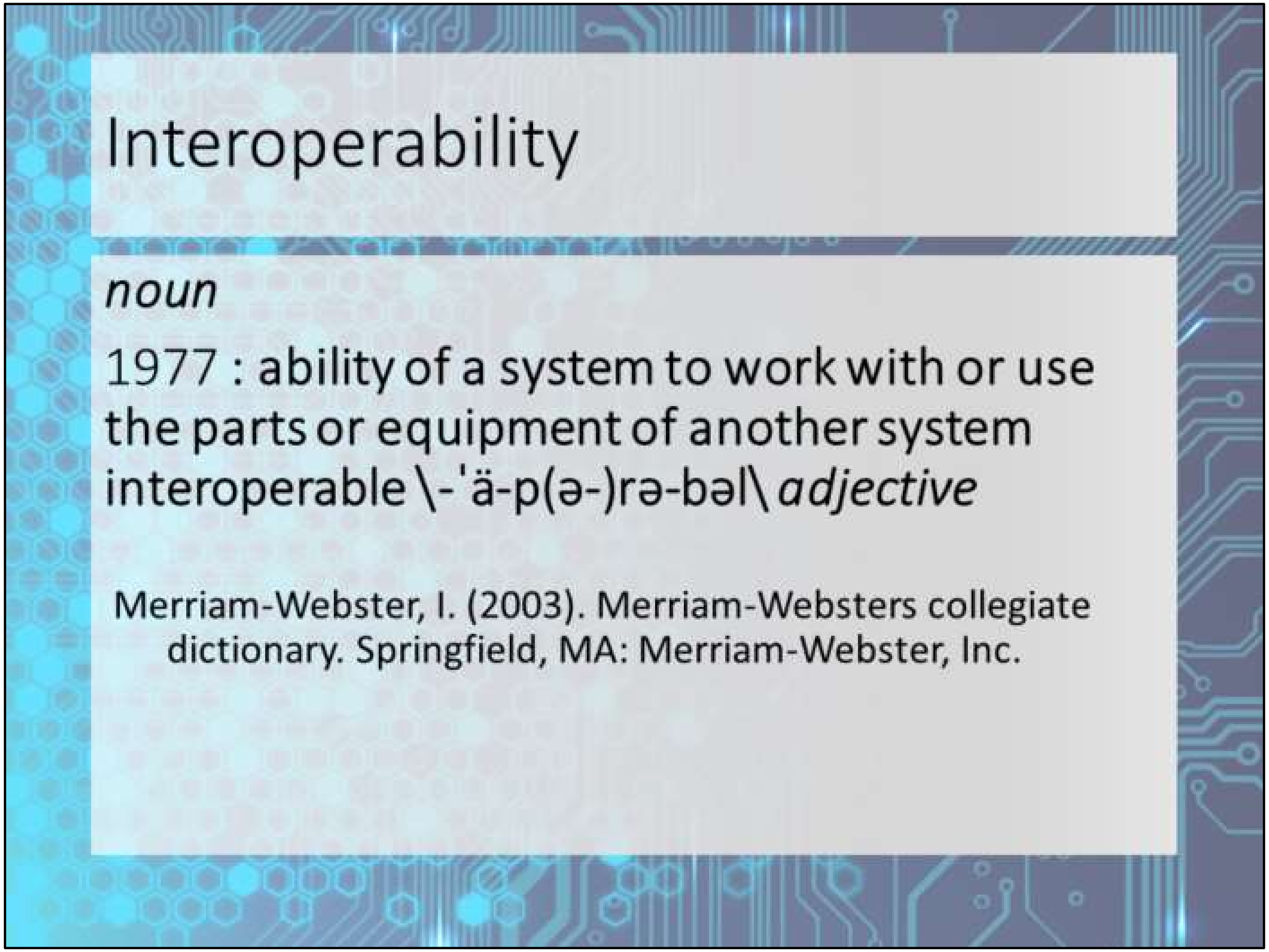
Chiarcos, Christian, Sebastian Nordhoff & Sebastian Hellman (eds.). 2012. *Linked Data in Linguistics*. Springer.

Full of articles detailing ways to make linguistic data smarter and to share, connect, and analyze it in new ways.



Course Goals:

- Learn about Linked Data and how it can be used in linguistics.
- Learn the technology necessary (XSLT) to transform XML Data.
- Transform XML lexical databases into an interoperable Linked Data format (RDF).
- Demonstrate interesting cross-linguistic searching across those databases.

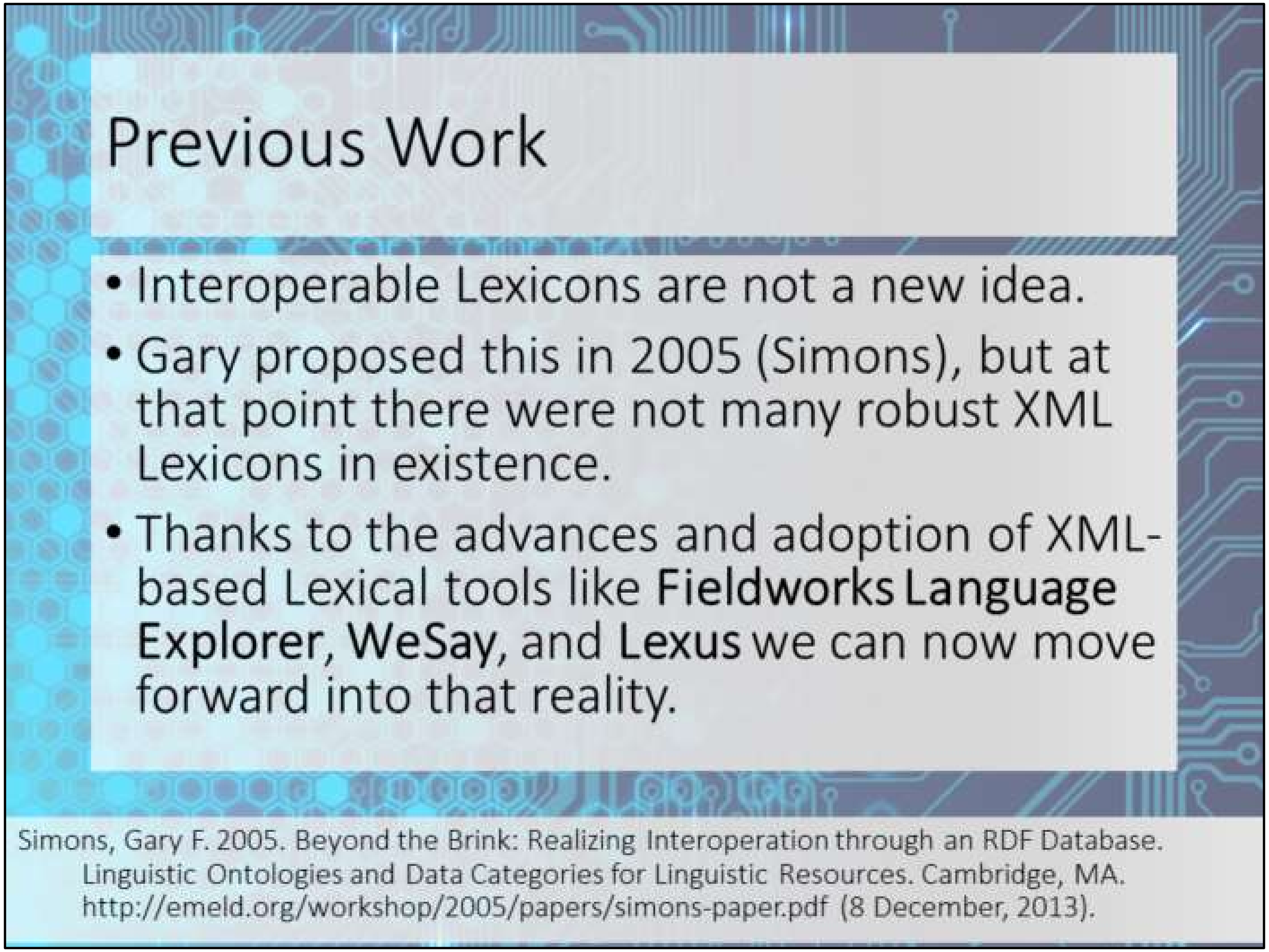


Interoperability

noun

1977 : ability of a system to work with or use the parts or equipment of another system interoperable \-ˈä-p(ə-)rə-bəl\ *adjective*

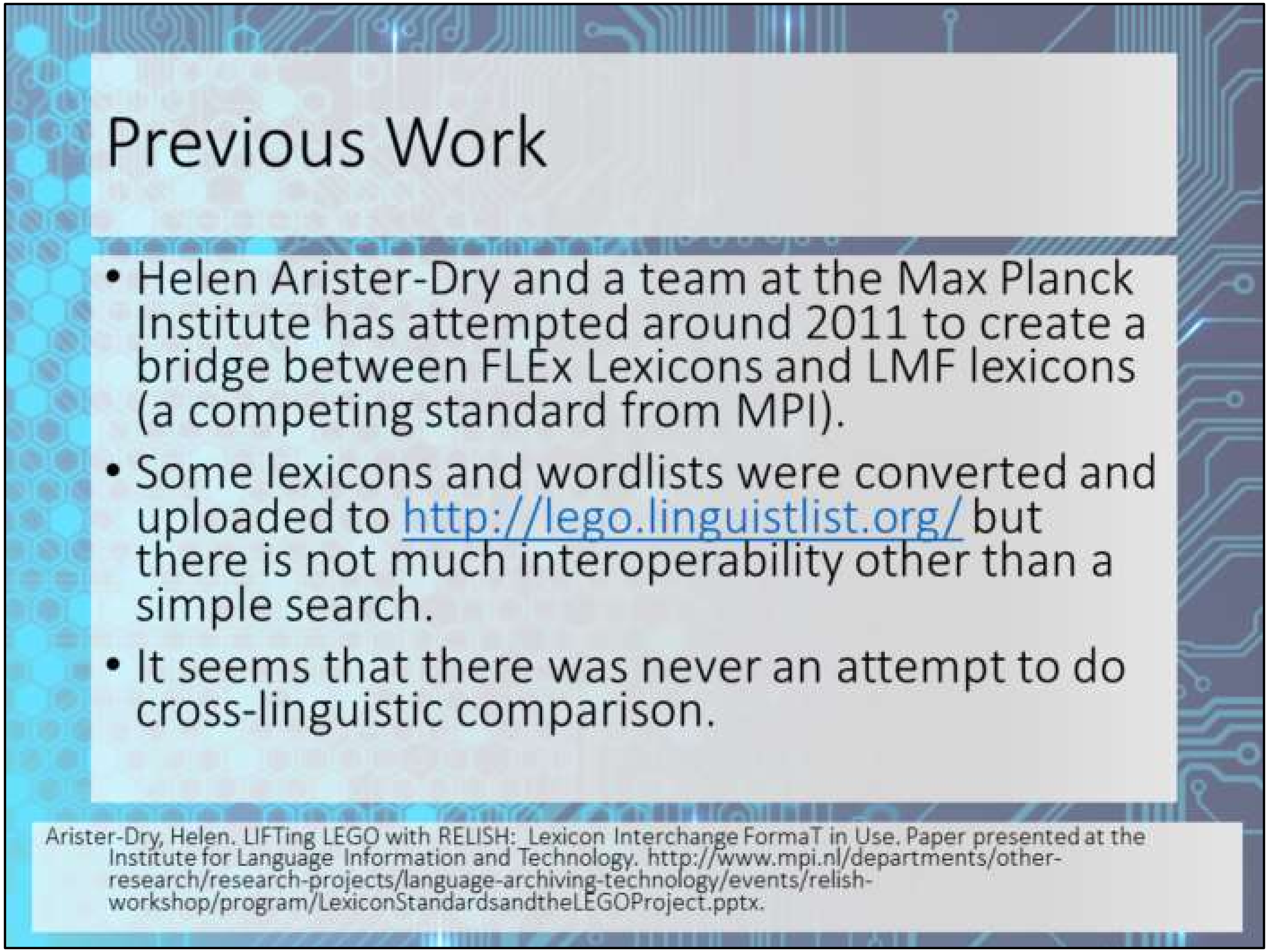
Merriam-Webster, I. (2003). Merriam-Websters collegiate dictionary. Springfield, MA: Merriam-Webster, Inc.



Previous Work

- Interoperable Lexicons are not a new idea.
- Gary proposed this in 2005 (Simons), but at that point there were not many robust XML Lexicons in existence.
- Thanks to the advances and adoption of XML-based Lexical tools like **Fieldworks Language Explorer**, **WeSay**, and **Lexus** we can now move forward into that reality.

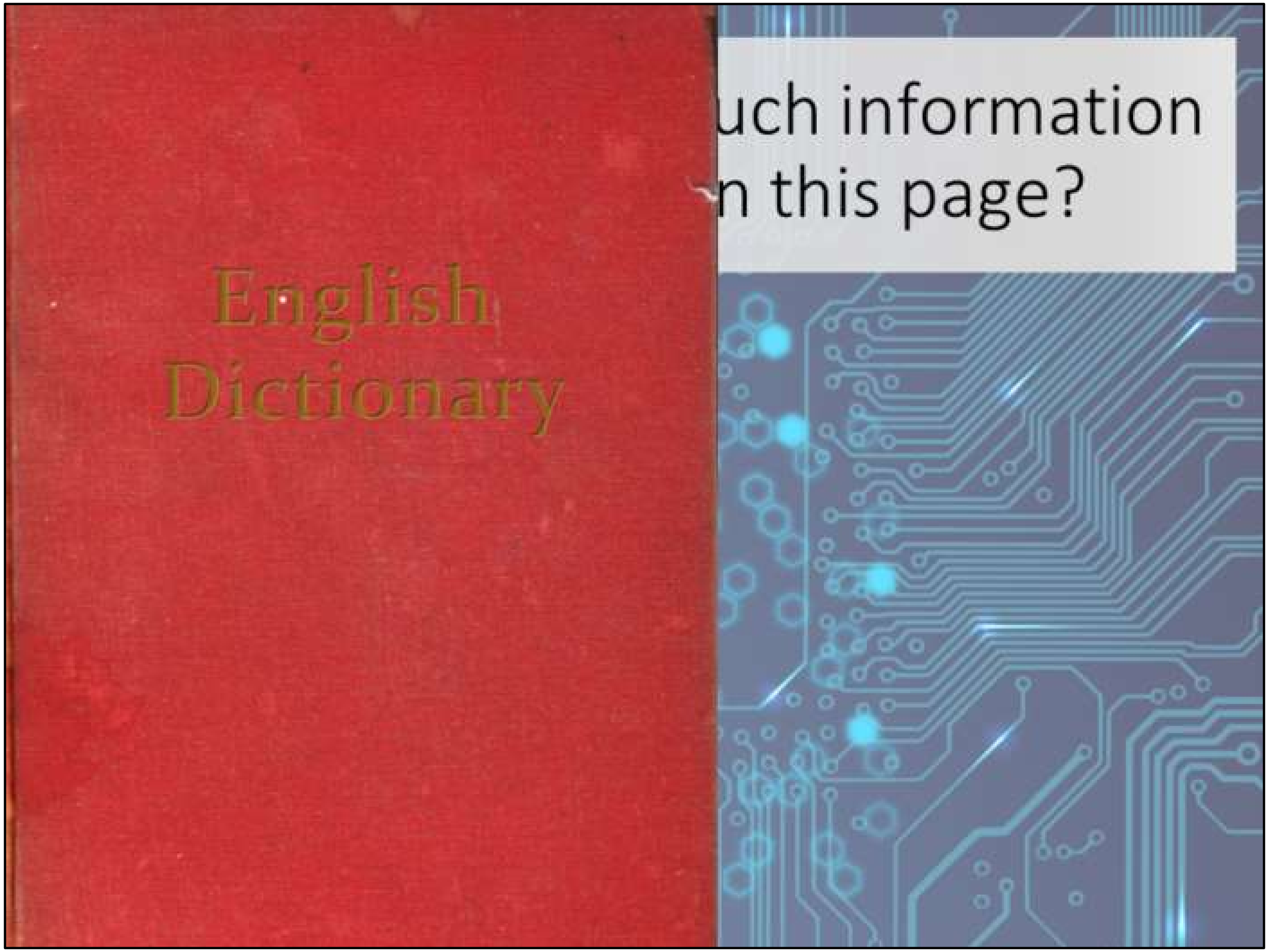
Simons, Gary F. 2005. Beyond the Brink: Realizing Interoperation through an RDF Database. Linguistic Ontologies and Data Categories for Linguistic Resources. Cambridge, MA. http://emeld.org/workshop/2005/papers/simons-paper.pdf (8 December, 2013).



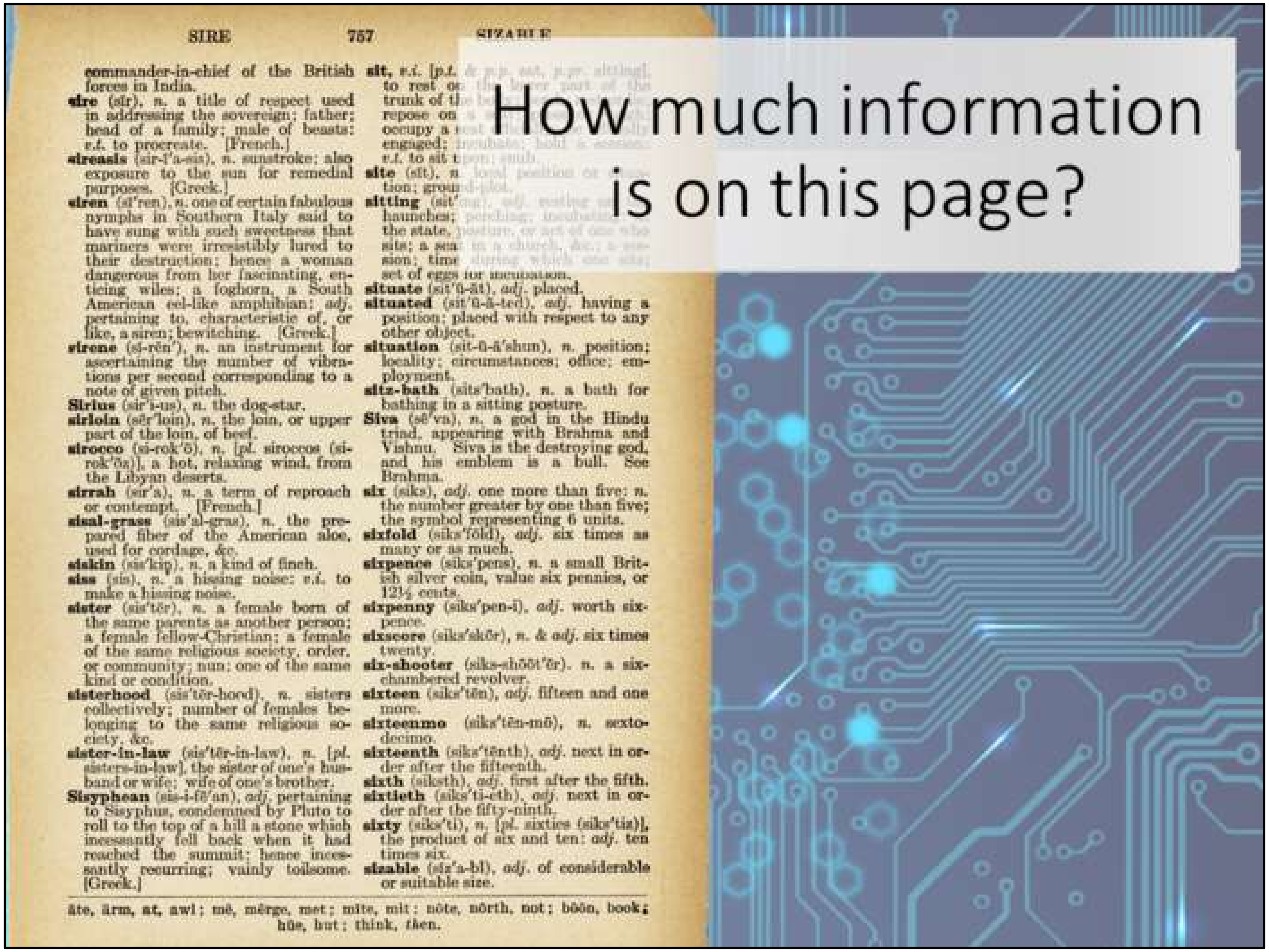
Previous Work

- Helen Arister-Dry and a team at the Max Planck Institute has attempted around 2011 to create a bridge between FLEx Lexicons and LMF lexicons (a competing standard from MPI).
- Some lexicons and wordlists were converted and uploaded to http://lego.linguistlist.org/ but there is not much interoperability other than a simple search.
- It seems that there was never an attempt to do cross-linguistic comparison.

Arister-Dry, Helen. LIFTing LEGO with RELISH: Lexicon Interchange FormaT in Use. Paper presented at the Institute for Language Information and Technology. http://www.mpi.nl/departments/other-research/research-projects/language-archiving-technology/events/relish-workshop/program/LexiconStandardsandtheLEGOProject.pptx.



uch information
n this page?



How much information is on this page?

SIRE 757 SIZABLE

commander-in-chief of the British forces in India.

sire (sīr), *n.* a title of respect used in addressing the sovereign; father; head of a family; male of beasts: *v.t.* to procreate. [French.]

siriasis (sir-ī'a-sis), *n.* sunstroke; also exposure to the sun for remedial purposes. [Greek.]

siren (sī'ren), *n.* one of certain fabulous nymphs in Southern Italy said to have sung with such sweetness that mariners were irresistibly lured to their destruction; hence a woman dangerous from her fascinating, enticing wiles; a foghorn, a South American eel-like amphibian: *adj.* pertaining to, characteristic of, or like, a siren; bewitching. [Greek.]

sirene (sī-rēn'), *n.* an instrument for ascertaining the number of vibrations per second corresponding to a note of given pitch.

Sirius (sir'i-us), *n.* the dog-star.

sirloin (sẽr'loin), *n.* the loin, or upper part of the loin, of beef.

sirocco (si-rok'ō), *n.* [*pl.* siroccos (si-rok'ōz)], a hot, relaxing wind, from the Libyan deserts.

sirrah (sir'a), *n.* a term of reproach or contempt. [French.]

sisal-grass (sis'al-gras), *n.* the prepared fiber of the American aloe, used for cordage, &c.

siskin (sis'kin), *n.* a kind of finch.

siss (sis), *n.* a hissing noise: *v.i.* to make a hissing noise.

sister (sis'tẽr), *n.* a female born of the same parents as another person; a female fellow-Christian; a female of the same religious society, order, or community; nun; one of the same kind or condition.

sisterhood (sis'tẽr-hood), *n.* sisters collectively; number of females belonging to the same religious society, &c.

sister-in-law (sis'tẽr-in-law), *n.* [*pl.* sisters-in-law], the sister of one's husband or wife; wife of one's brother.

Sisyphean (sis-i-fē'an), *adj.* pertaining to Sisyphus, condemned by Pluto to roll to the top of a hill a stone which incessantly fell back when it had reached the summit; hence incessantly recurring; vainly toilsome. [Greek.]

sit, *v.i.* [*p.t.* & *p.p.* sat, *p.pr.* sitting], to rest on the lower part of the trunk of the b[illegible] repose on a [illegible] occupy a seat [illegible] engaged; incubate; hold a session: *v.t.* to sit upon; snub.

site (sīt), *n.* local position or situation; ground-plot.

sitting (sit'ing), *adj.* resting on haunches; perching; incubating: *n.* the state, posture, or act of one who sits; a seat in a church, &c.; a session; time during which one sits; set of eggs for incubation.

situate (sit'ū-āt), *adj.* placed.

situated (sit'ū-ā-ted), *adj.* having a position; placed with respect to any other object.

situation (sit-ū-ā'shun), *n.* position; locality; circumstances; office; employment.

sitz-bath (sits'bath), *n.* a bath for bathing in a sitting posture.

Siva (sē'va), *n.* a god in the Hindu triad, appearing with Brahma and Vishnu. Siva is the destroying god, and his emblem is a bull. See Brahma.

six (siks), *adj.* one more than five: *n.* the number greater by one than five; the symbol representing 6 units.

sixfold (siks'fōld), *adj.* six times as many or as much.

sixpence (siks'pens), *n.* a small British silver coin, value six pennies, or 12½ cents.

sixpenny (siks'pen-i), *adj.* worth sixpence.

sixscore (siks'skōr), *n.* & *adj.* six times twenty.

six-shooter (siks-shōōt'ẽr), *n.* a six-chambered revolver.

sixteen (siks'tēn), *adj.* fifteen and one more.

sixteenmo (siks'tēn-mō), *n.* sextodecimo.

sixteenth (siks'tēnth), *adj.* next in order after the fifteenth.

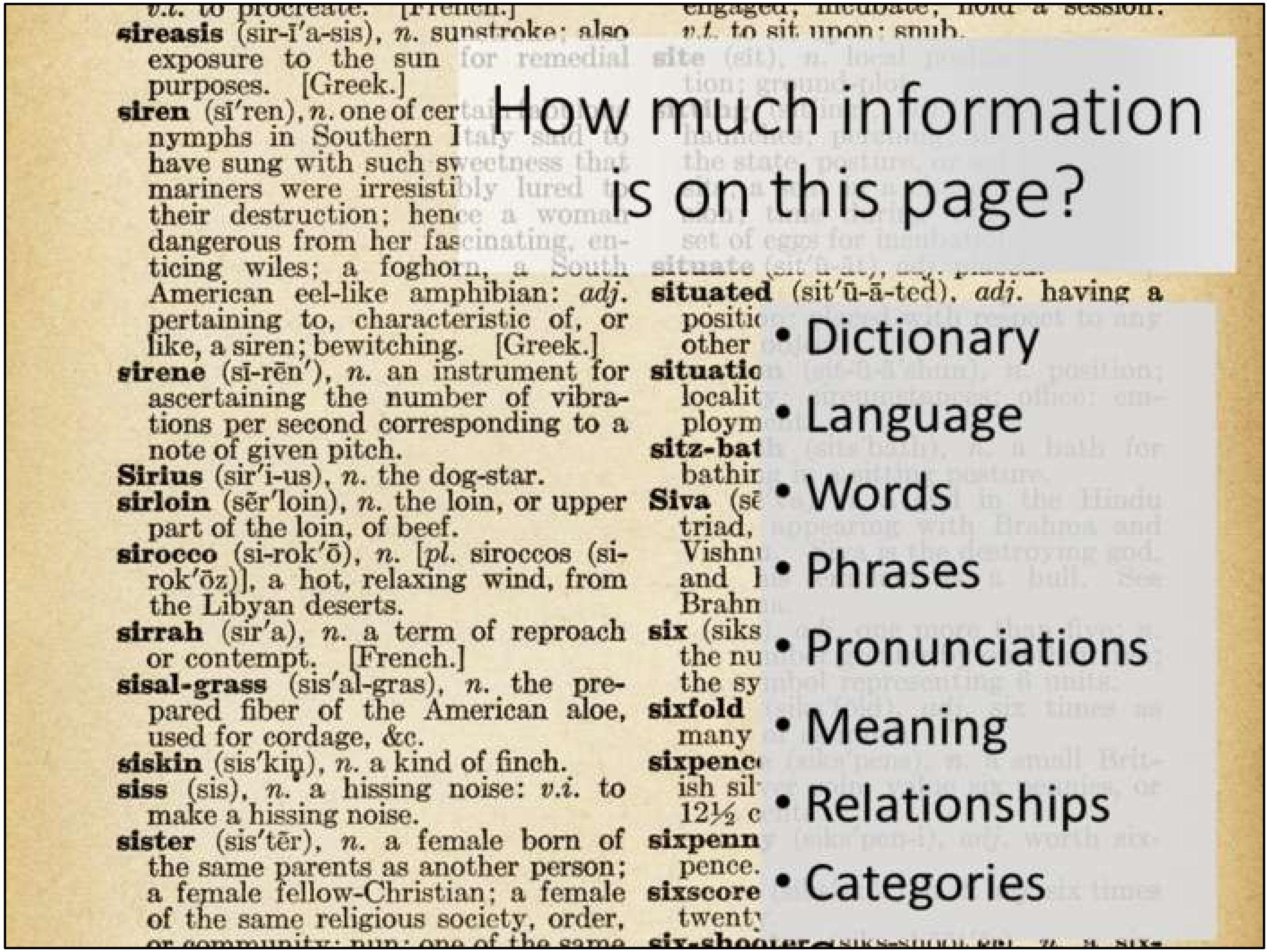
sixth (siksth), *adj.* first after the fifth.

sixtieth (siks'ti-eth), *adj.* next in order after the fifty-ninth.

sixty (siks'ti), *n.* [*pl.* sixties (siks'tiz)], the product of six and ten: *adj.* ten times six.

sizable (sīz'a-bl), *adj.* of considerable or suitable size.

āte, ärm, at, awl; mē, mẽrge, met; mīte, mit; nōte, nôrth, not; bōōn, book; hūe, hut; think, *then.*



How much information is on this page?

- Dictionary
- Language
- Words
- Phrases
- Pronunciations
- Meaning
- Relationships
- Categories

v.i. to procreate. [French.]

sireasis (sir-ī′a-sis), *n.* sunstroke; also exposure to the sun for remedial purposes. [Greek.]

siren (sī′ren), *n.* one of certain nymphs in Southern Italy said to have sung with such sweetness that mariners were irresistibly lured to their destruction; hence a woman dangerous from her fascinating, enticing wiles; a foghorn, a South American eel-like amphibian: *adj.* pertaining to, characteristic of, or like, a siren; bewitching. [Greek.]

sirene (sī-rēn′), *n.* an instrument for ascertaining the number of vibrations per second corresponding to a note of given pitch.

Sirius (sir′i-us), *n.* the dog-star.

sirloin (sẽr′loin), *n.* the loin, or upper part of the loin, of beef.

sirocco (si-rok′ō), *n.* [*pl.* siroccos (si-rok′ōz)], a hot, relaxing wind, from the Libyan deserts.

sirrah (sir′a), *n.* a term of reproach or contempt. [French.]

sisal-grass (sis′al-gras), *n.* the prepared fiber of the American aloe, used for cordage, &c.

siskin (sis′kin), *n.* a kind of finch.

siss (sis), *n.* a hissing noise: *v.i.* to make a hissing noise.

sister (sis′tẽr), *n.* a female born of the same parents as another person; a female fellow-Christian; a female of the same religious society, order, or community; nun; one of the same

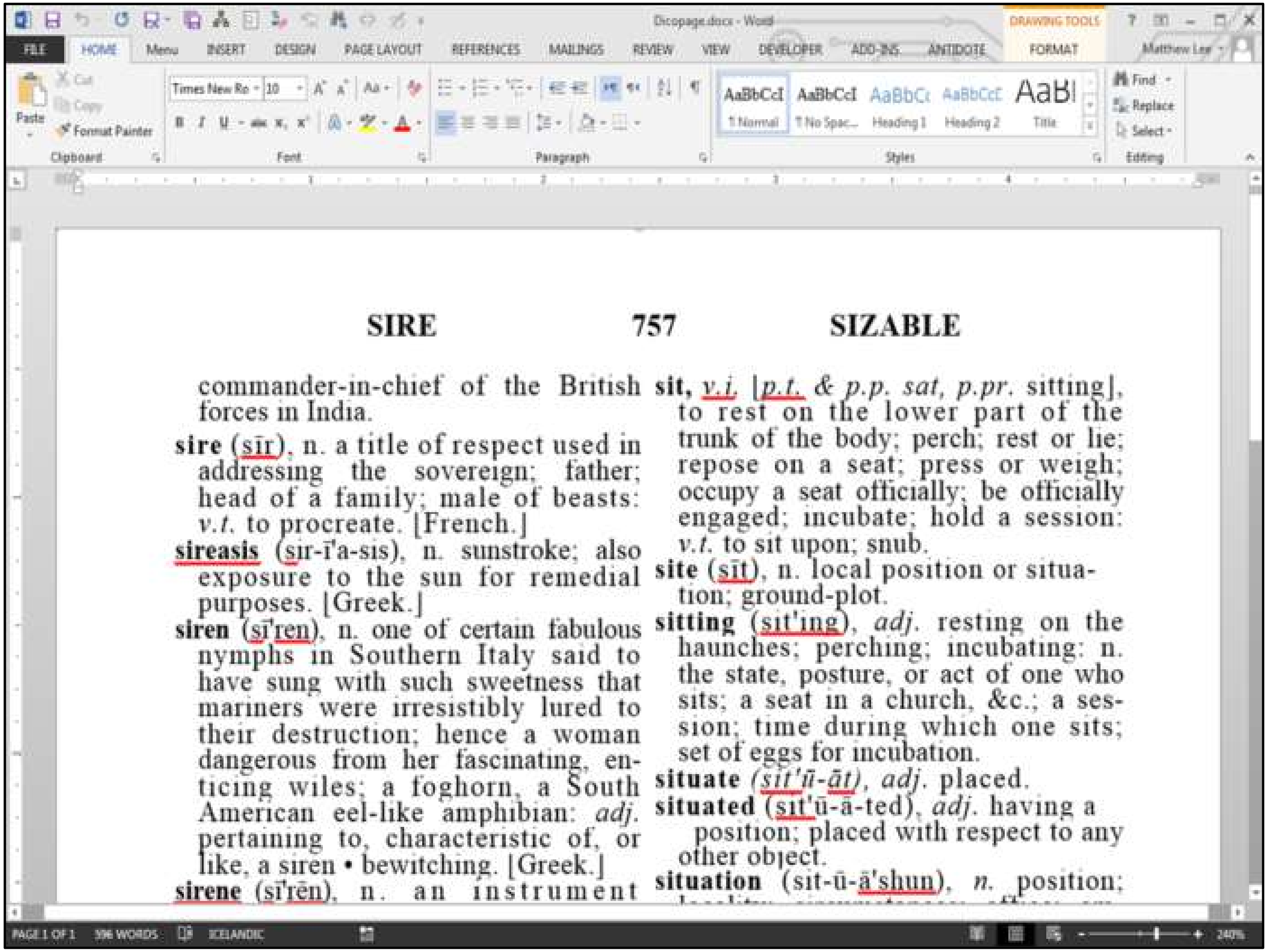
engaged; incubate; hold a session: *v.t.* to sit upon; snub.

site (sīt), *n.* local position; ground-plot.

situated (sit′ū-ā-ted), *adj.* having a position

Your brain knows how to interpret this data, it sees: [Click]

Dictionaries as Datapoints: Matthew Lee



SIRE **757** **SIZABLE**

commander-in-chief of the British forces in India.

sire (sīr), n. a title of respect used in addressing the sovereign; father; head of a family; male of beasts: *v.t.* to procreate. [French.]

sireasis (sir-ī'a-sis), n. sunstroke; also exposure to the sun for remedial purposes. [Greek.]

siren (sī'ren), n. one of certain fabulous nymphs in Southern Italy said to have sung with such sweetness that mariners were irresistibly lured to their destruction; hence a woman dangerous from her fascinating, enticing wiles; a foghorn, a South American eel-like amphibian: *adj.* pertaining to, characteristic of, or like, a siren • bewitching. [Greek.]

sirene (sī'rēn), n. an instrument

sit, *v.i.* [*p.t. & p.p. sat, p.pr.* sitting], to rest on the lower part of the trunk of the body; perch; rest or lie; repose on a seat; press or weigh; occupy a seat officially; be officially engaged; incubate; hold a session: *v.t.* to sit upon; snub.

site (sīt), n. local position or situation; ground-plot.

sitting (sit'ing), *adj.* resting on the haunches; perching; incubating: n. the state, posture, or act of one who sits; a seat in a church, &c.; a session; time during which one sits; set of eggs for incubation.

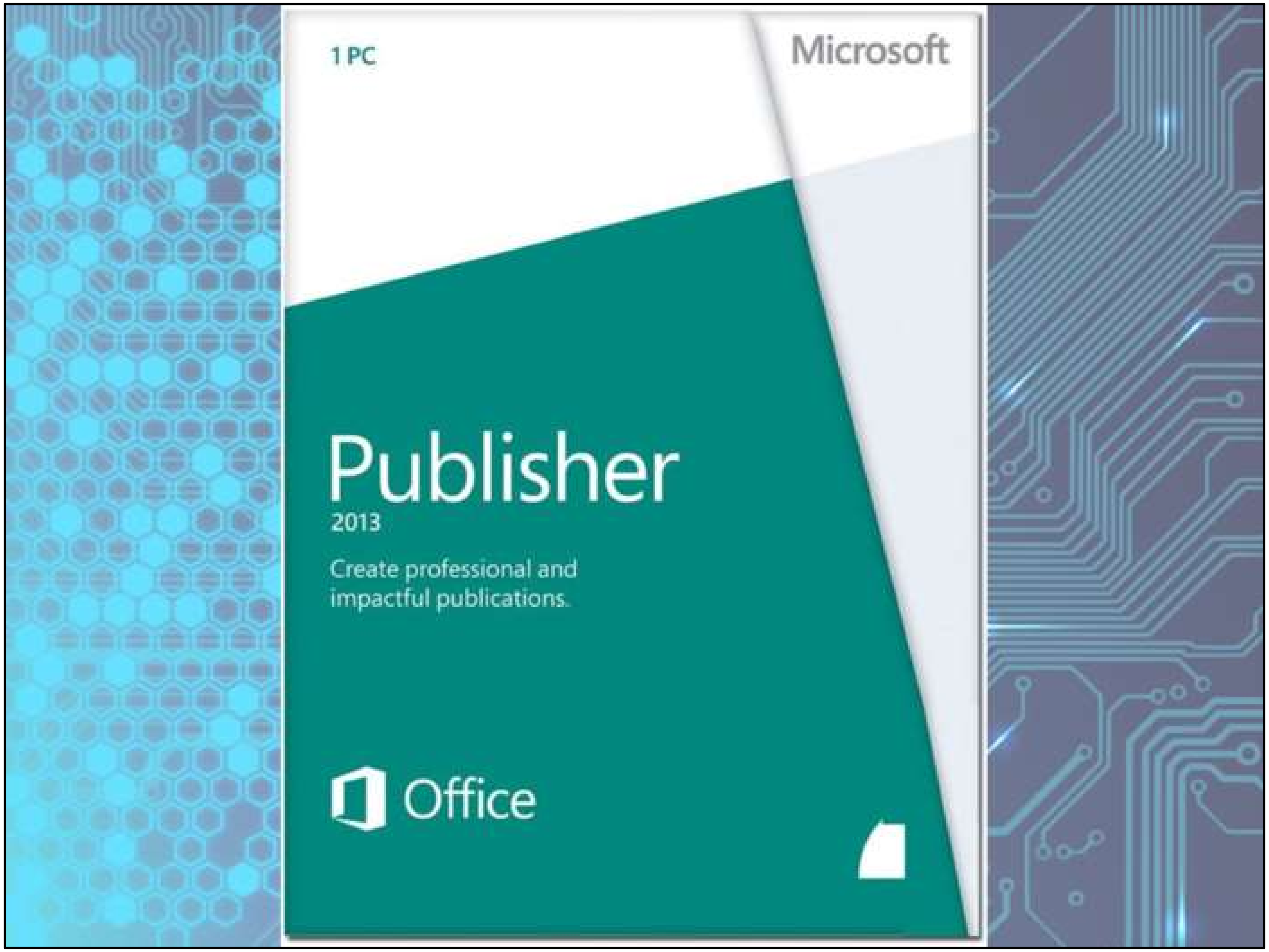
situate *(sit'ū-āt), adj.* placed.

situated (sit'ū-ā-ted), *adj.* having a position; placed with respect to any other object.

situation (sit-ū-ā'shun), *n.* position;

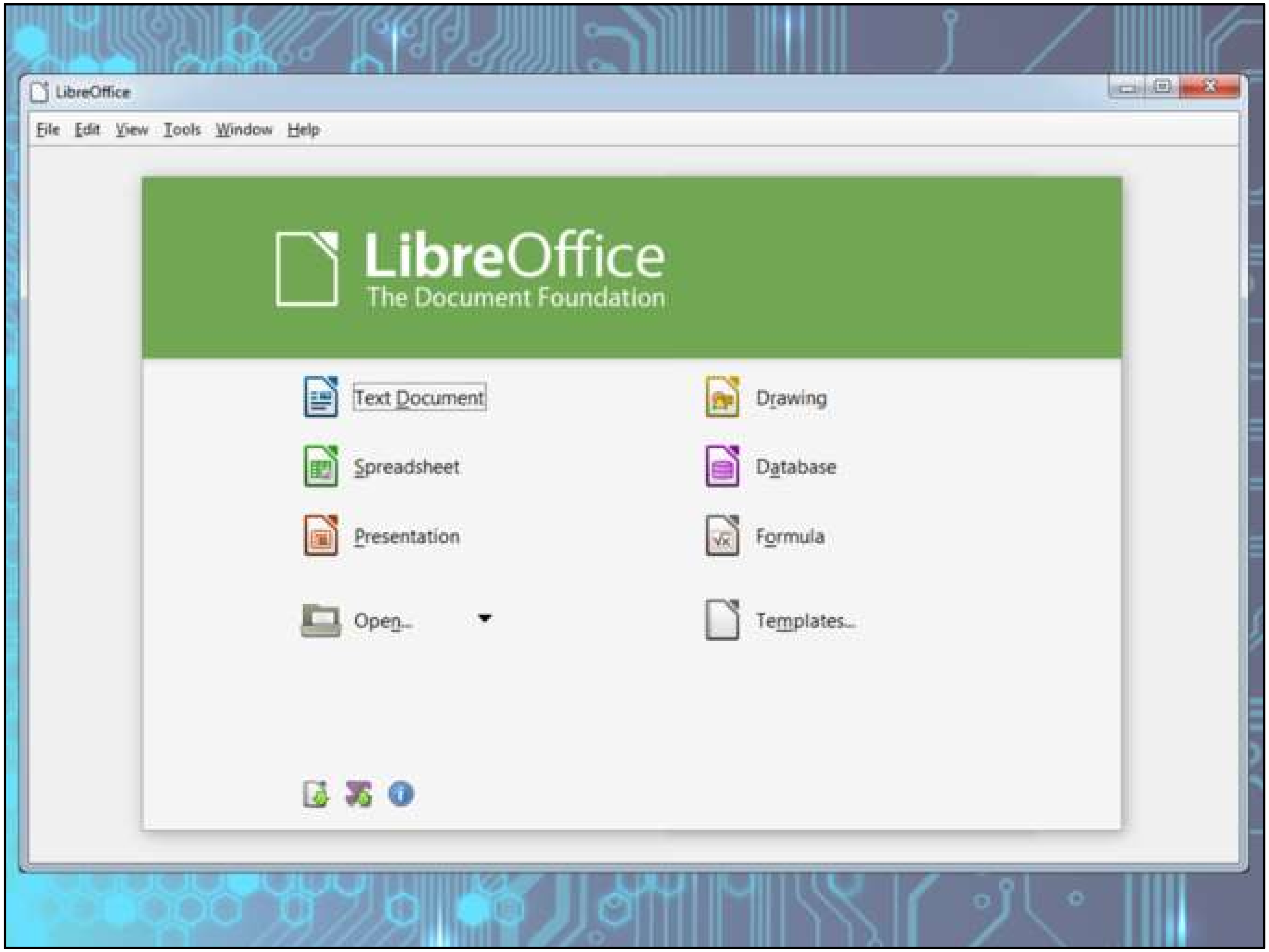
Let’s move this information into a Desktop Publishing program….for example Microsoft Word.
Just so you know I’m not picking on word…this could be [Click]

Dictionaries as Datapoints: Matthew Lee



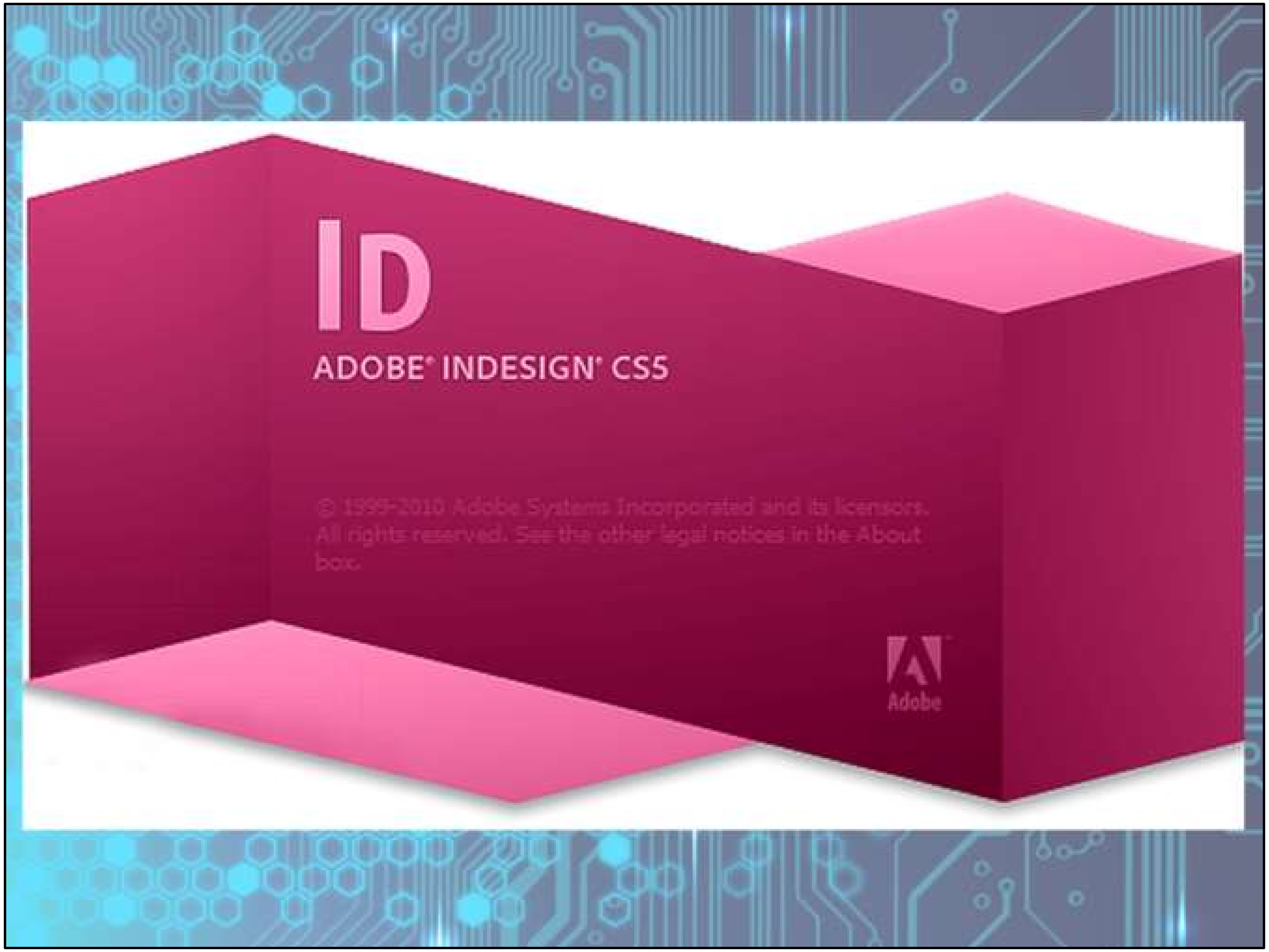
Microsoft Publisher

Dictionaries as Datapoints: Matthew Lee



Libreoffice or OpenOffice

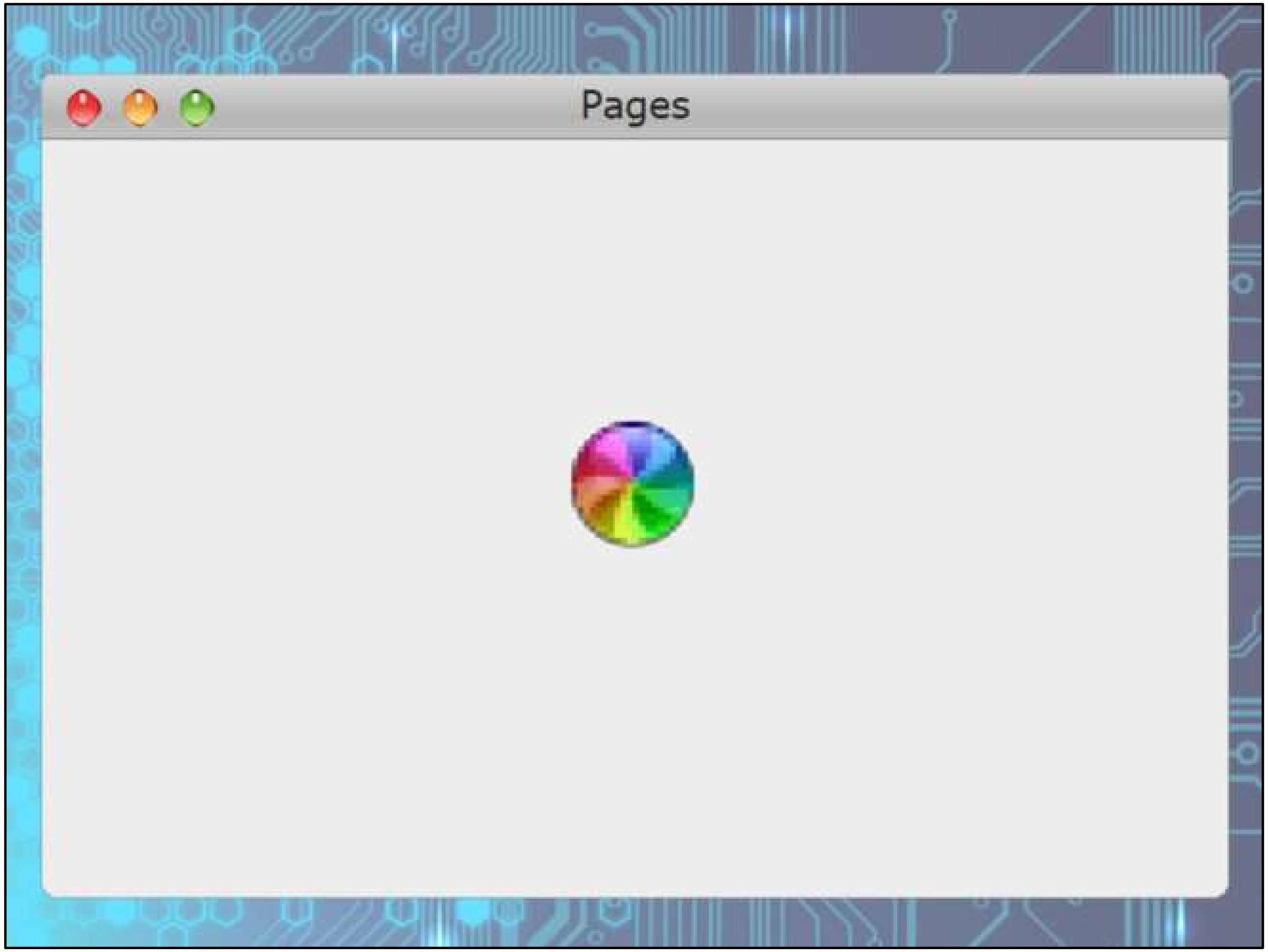
Dictionaries as Datapoints: Matthew Lee



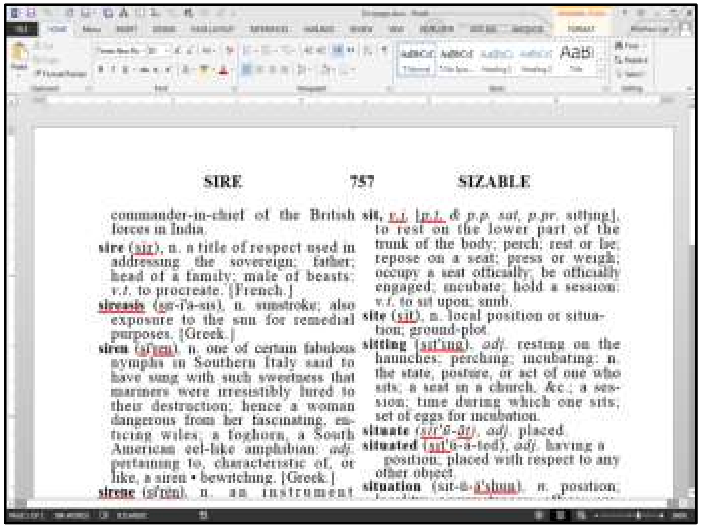
© 1999-2010 Adobe Systems Incorporated and its licensors. All rights reserved. See the other legal notices in the About box.

Adobe InDesign

Dictionaries as Datapoints: Matthew Lee



Or Apple’s Pages



SIRE 757 SIZABLE

commander-in-chief of the British forces in India.

sire (sīr), n. a title of respect used in addressing the sovereign; father; head of a family; male of beasts: *v.t.* to procreate. [French.]

sireasis (sĭr-ī'a-sis), n. sunstroke; also exposure to the sun for remedial purposes. [Greek.]

siren (sī'ren), n. one of certain fabulous nymphs in Southern Italy said to have sung with such sweetness that mariners were irresistibly lured to their destruction; hence a woman dangerous from her fascinating, enticing wiles; a foghorn, a South American eel-like amphibian: *adj.* pertaining to, characteristic of, or like, a siren • bewitching. [Greek.]

sirene (sī'rēn), n. an instrument

sit, *v.i.* [*p.t.* & *p.p. sat, p.pr.* sitting], to rest on the lower part of the trunk of the body; perch; rest or lie; repose on a seat; press or weigh; occupy a seat officially; be officially engaged; incubate; hold a session: *v.t.* to sit upon; snub.

site (sīt), n. local position or situation; ground-plot.

sitting (sit'ing), *adj.* resting on the haunches; perching; incubating: n. the state, posture, or act of one who sits; a seat in a church, &c.; a session; time during which one sits; set of eggs for incubation.

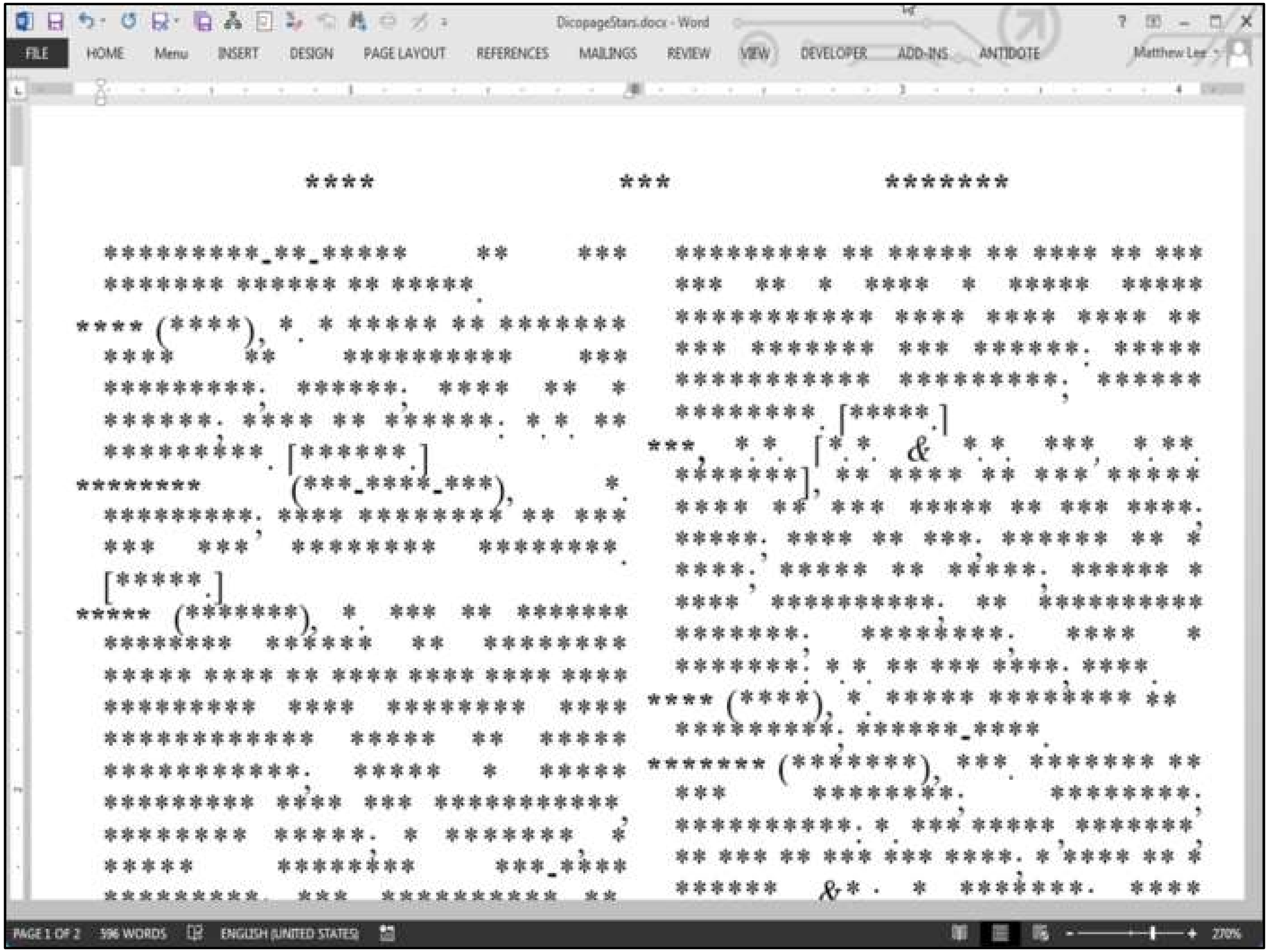
situate (*sit'ū-āt*), *adj.* placed.

situated (sit'ū-ā-ted), *adj.* having a position; placed with respect to any other object.

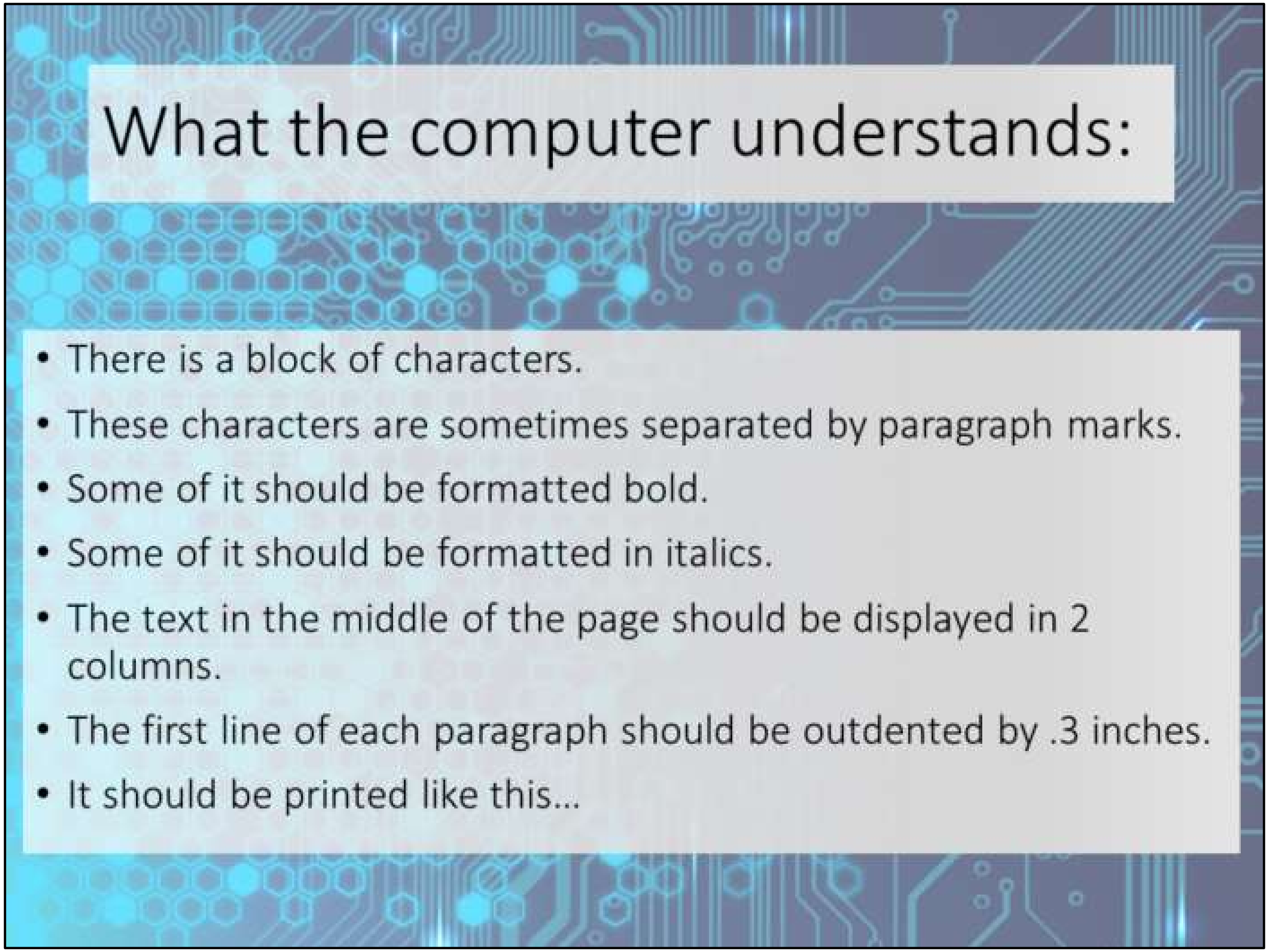
situation (sit-ū-ā'shun), *n.* position;

Back to Word…these programs are designed for creating formatted documents, they lack critical features for organizing data. What does Word “understand” about this document?

Dictionaries as Datapoints: Matthew Lee



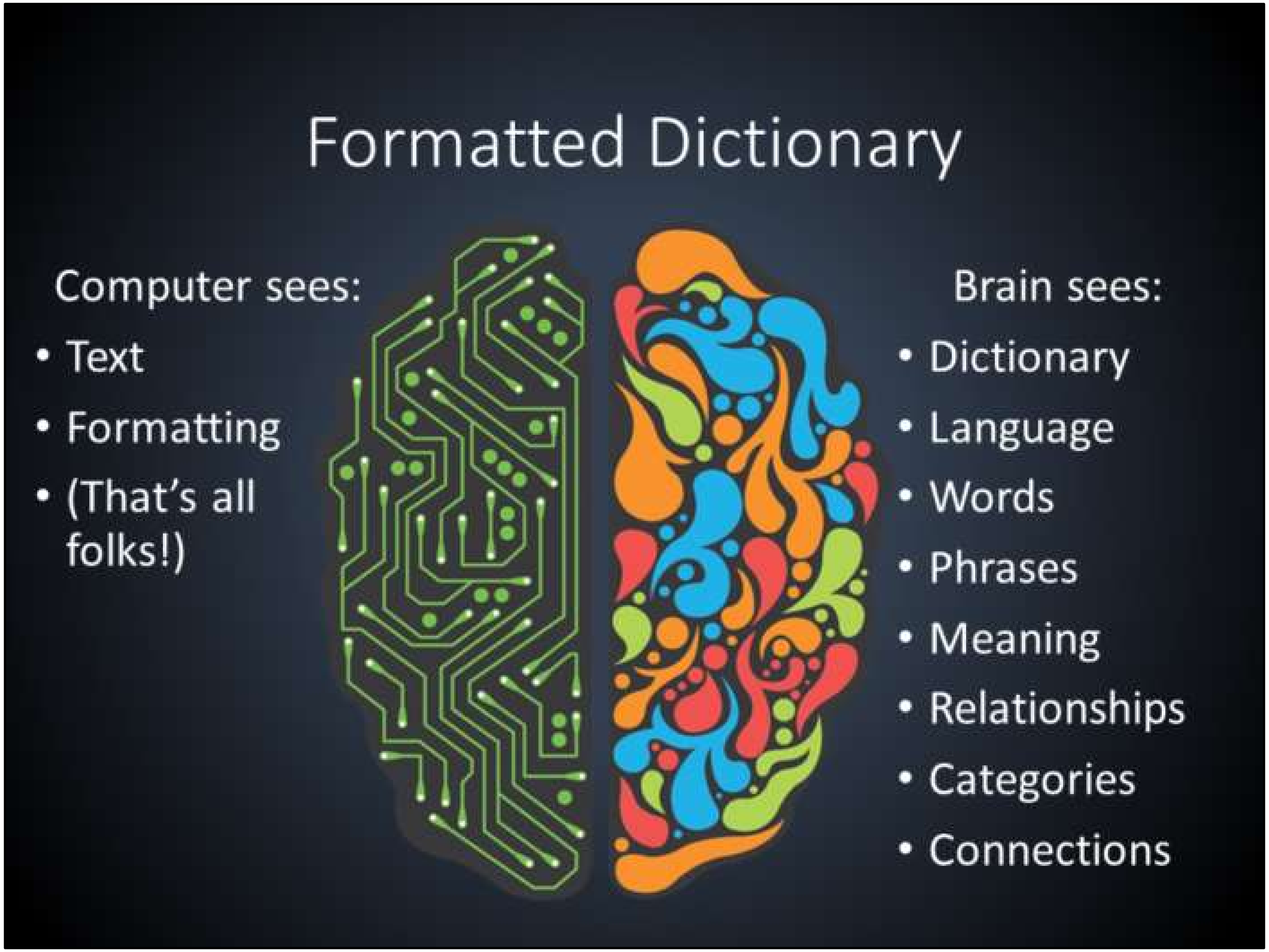
Though it can reliably print you document, Word knows almost nothing about the content of a document. How many facts can the computer access? Can you ask it for a list of lexemes? Can Word help you to create a reversal for your dictionary?



What the computer understands:

- There is a block of characters.
- These characters are sometimes separated by paragraph marks.
- Some of it should be formatted bold.
- Some of it should be formatted in italics.
- The text in the middle of the page should be displayed in 2 columns.
- The first line of each paragraph should be outdented by .3 inches.
- It should be printed like this…

This is the reason why…if you want to create a document in another form, for printing on A4 or a different audience…you now have two documents and you must be careful to make consistent changes to both.



Formatted Dictionary

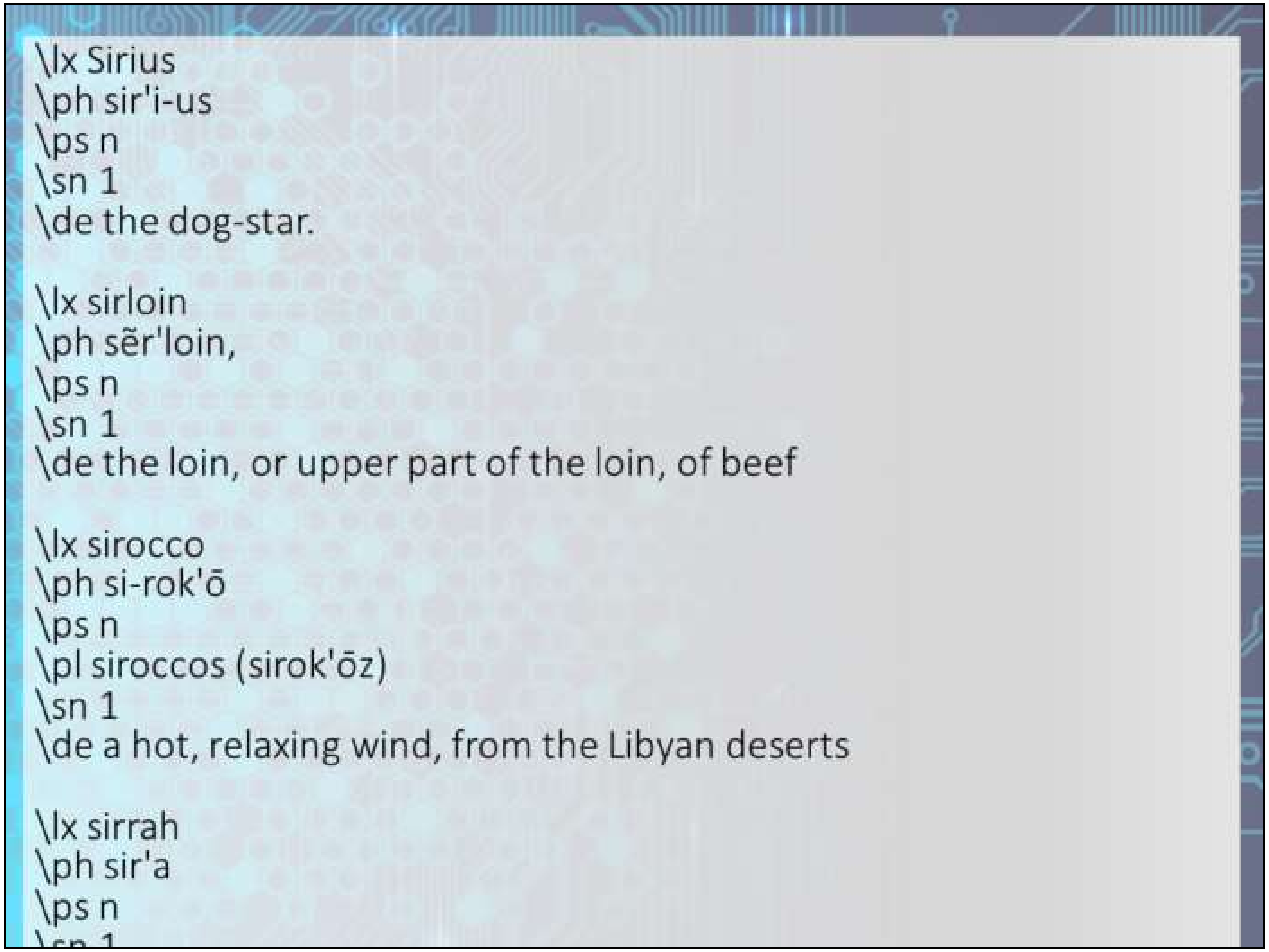
Computer sees:

- Text
- Formatting
- (That's all folks!)

Brain sees:

- Dictionary
- Language
- Words
- Phrases
- Meaning
- Relationships
- Categories
- Connections

Lexical Databases: A Step Forward



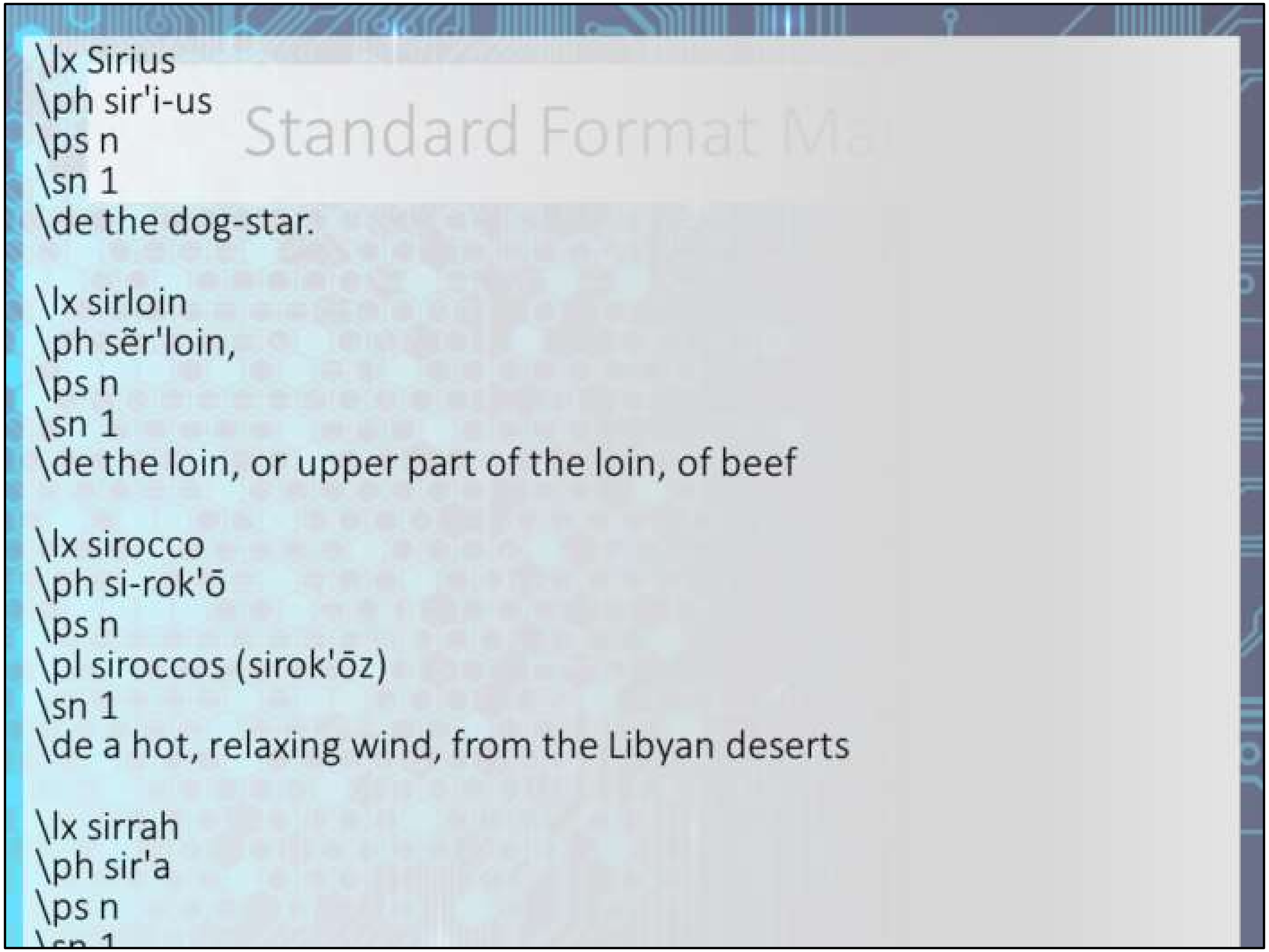
```
\lx Sirius
\ph sir'i-us
\ps n
\sn 1
\de the dog-star.

\lx sirloin
\ph sẽr'loin,
\ps n
\sn 1
\de the loin, or upper part of the loin, of beef

\lx sirocco
\ph si-rok'ō
\ps n
\pl siroccos (sirok'ōz)
\sn 1
\de a hot, relaxing wind, from the Libyan deserts

\lx sirrah
\ph sir'a
\ps n
\sn 1
```

I have converted the previous document into another format. How many of you recognize this lexicon format?



Standard Format Ma

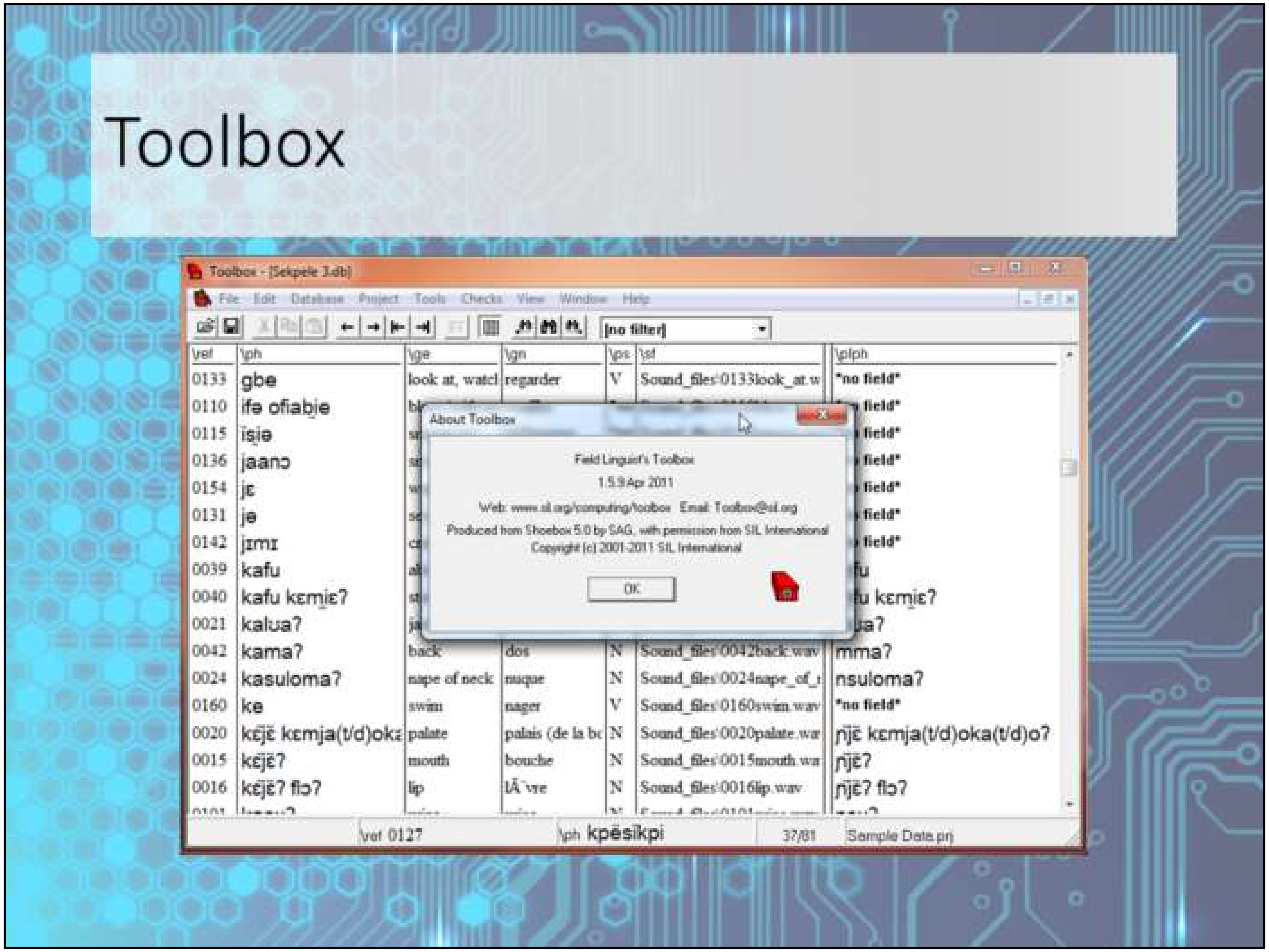
```
\lx Sirius
\ph sir'i-us
\ps n
\sn 1
\de the dog-star.

\lx sirloin
\ph sẽr'loin,
\ps n
\sn 1
\de the loin, or upper part of the loin, of beef

\lx sirocco
\ph si-rok'ō
\ps n
\pl siroccos (sirok'ōz)
\sn 1
\de a hot, relaxing wind, from the Libyan deserts

\lx sirrah
\ph sir'a
\ps n
\sn 1
```

This is the Standard Format, the file used by Shoebox and Toolbox. I count 282 “Facts”. Now I can search and filter the data if I put it into Toolbox.



Toolbox

\ref	\ph	\ge	\gn	\ps	\sf	\plph
0133	gbe	look at, watcl	regarder	V	Sound_files\0133look_at.w	*no field*
0110	ifə ofiabi̯e	bl				field*
0115	ĩsi̯ə	sr				field*
0136	jaanɔ	sr				field*
0154	jɛ	w				field*
0131	jə	se				field*
0142	jɪmɪ	cr				field*
0039	kafu	al				fu
0040	kafu kɛmi̯ɛʔ	st				fu kɛmi̯ɛʔ
0021	kaluaʔ	ja				aʔ
0042	kamaʔ	back	dos	N	Sound_files\0042back.wav	mmaʔ
0024	kasulomaʔ	nape of neck	naque	N	Sound_files\0024nape_of_r	nsulomaʔ
0160	ke	swim	nager	V	Sound_files\0160swim.wav	*no field*
0020	kɛ̃jɛ̃ kɛmja(t/d)oka	palate	palais (de la bc	N	Sound_files\0020palate.wa	ɲɛ̃ kɛmja(t/d)oka(t/d)oʔ
0015	kɛ̃jɛ̃ʔ	mouth	bouche	N	Sound_files\0015mouth.wa	ɲɛ̃ʔ
0016	kɛ̃jɛ̃ʔ flɔʔ	lip	lÃ¨vre	N	Sound_files\0016lip.wav	ɲɛ̃ʔ flɔʔ

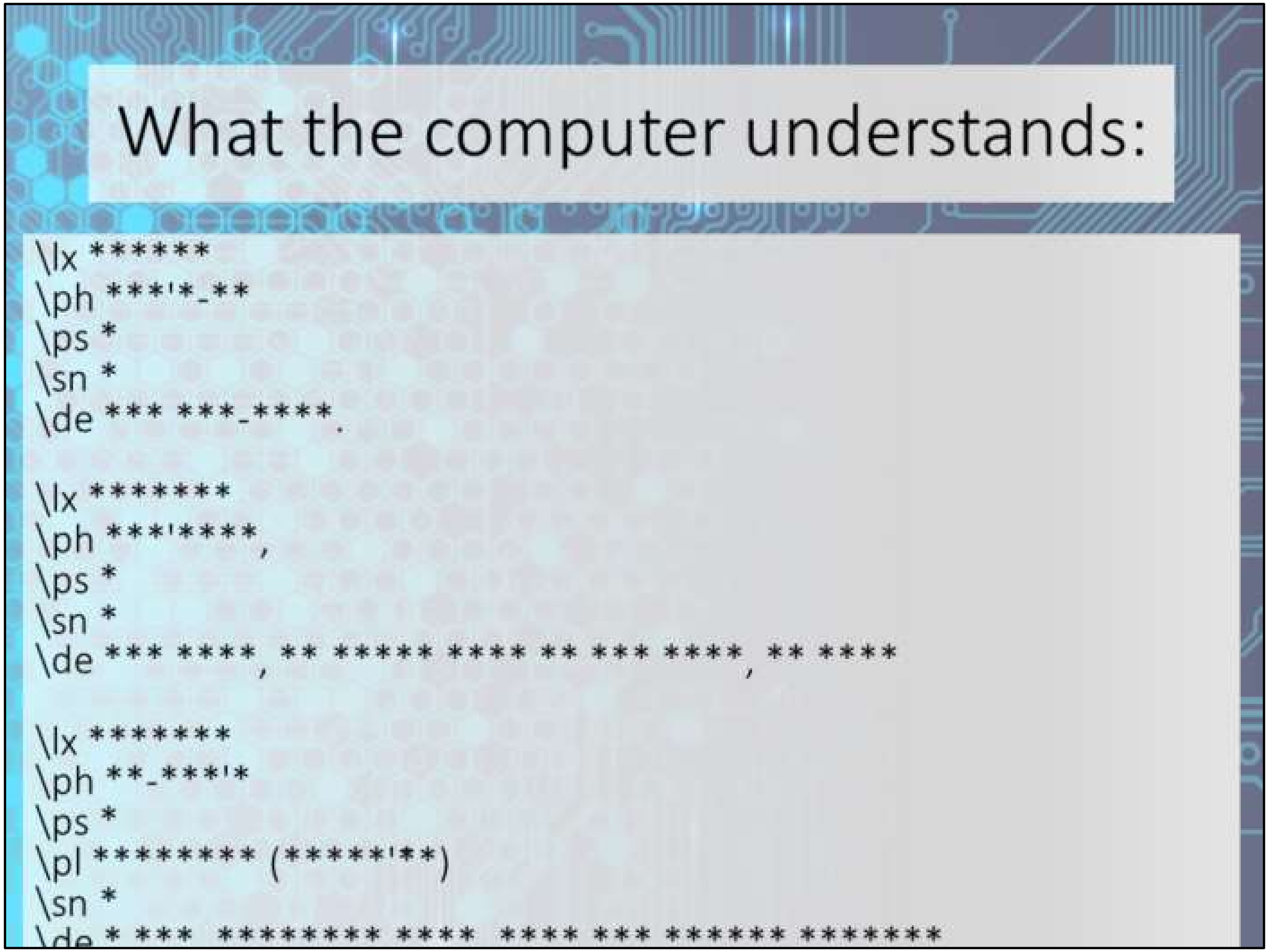
About Toolbox

Field Linguist's Toolbox
1.5.9 Apr 2011
Web: www.sil.org/computing/toolbox Email: Toolbox@sil.org
Produced from Shoebox 5.0 by SAG, with permission from SIL International
Copyright (c) 2001-2011 SIL International

OK

\ref 0127 \ph kpësikpi 37/81 Sample Data.prj

People love Toolbox! The best part of Toolbox (for working) is that it doesn't constrain you to any consistent structure. The worst part of toolbox (for publishing) it that doesn't constrain you to any consistent structure. If you want to add a custom column that keeps track of which words you've taught your pet parrot…it will let you put that anywhere. As a result, Toolbox Lexicons tend to mix and match languages, abbreviations, order, and notations over the years. They also tend to have references to non-existent entries. There is no safety net.



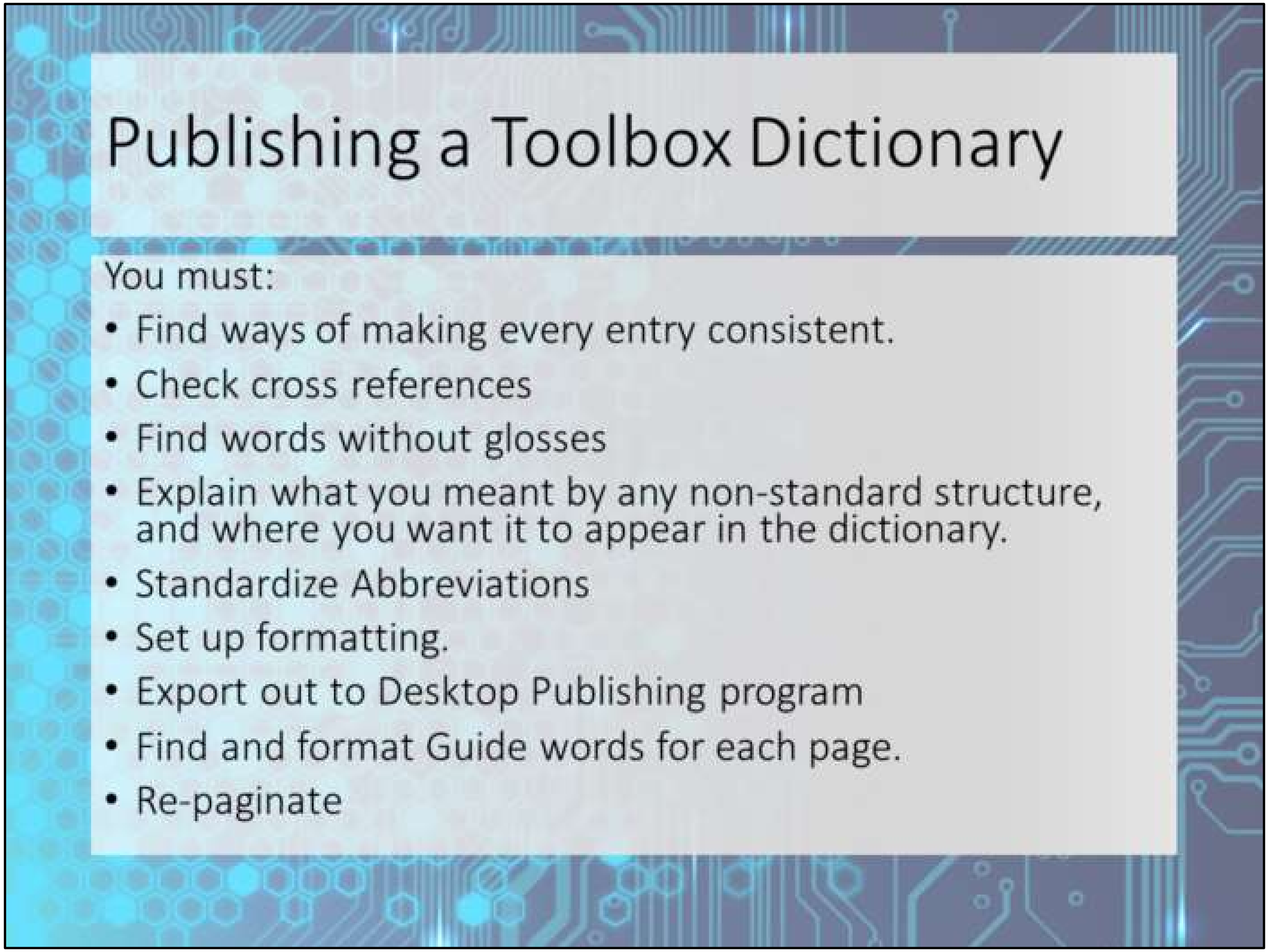
What the computer understands:

```
\lx ******
\ph ***'*_**
\ps *
\sn *
\de *** ***_****.

\lx *******
\ph ***'****,
\ps *
\sn *
\de *** ****, ** ***** **** ** *** ****, ** ****

\lx *******
\ph **_***'*
\ps *
\pl ******** (*****'**)
\sn *
\de * *** ******** **** **** *** ****** *******
```

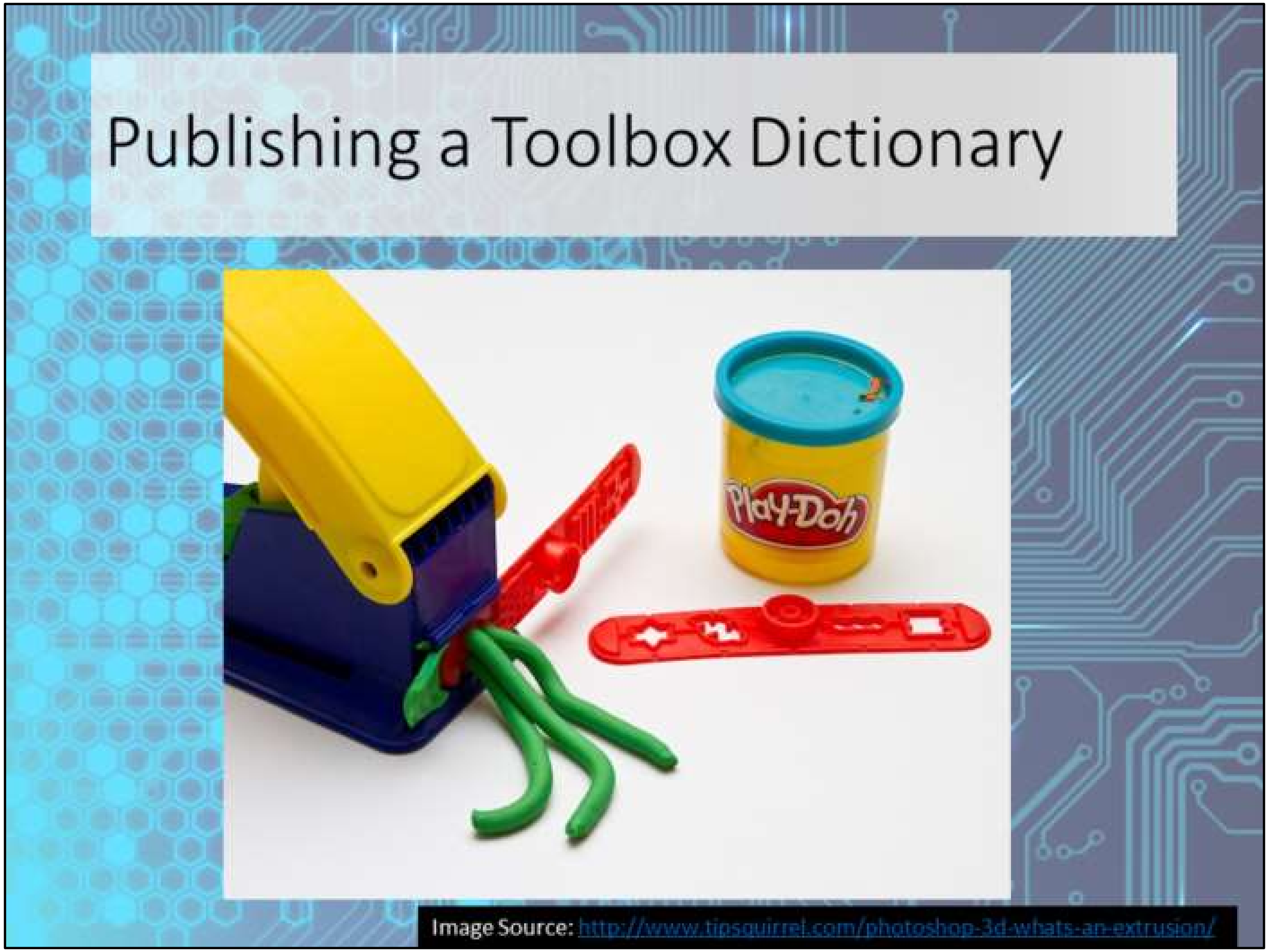
IN SFM, the content is still a mystery to the computer, but at least the program “knows” that each line represents a specific type of data. With the right configuration, Toolbox is designed to keep track of structured data in many languages. Though there are published standards, the structure here is really only understood by convention (Lexeme, part of speech, etc…), and you often need the original author to explain what some obscure items mean.



Publishing a Toolbox Dictionary

You must:

- Find ways of making every entry consistent.
- Check cross references
- Find words without glosses
- Explain what you meant by any non-standard structure, and where you want it to appear in the dictionary.
- Standardize Abbreviations
- Set up formatting.
- Export out to Desktop Publishing program
- Find and format Guide words for each page.
- Re-paginate



Publishing a Toolbox Dictionary

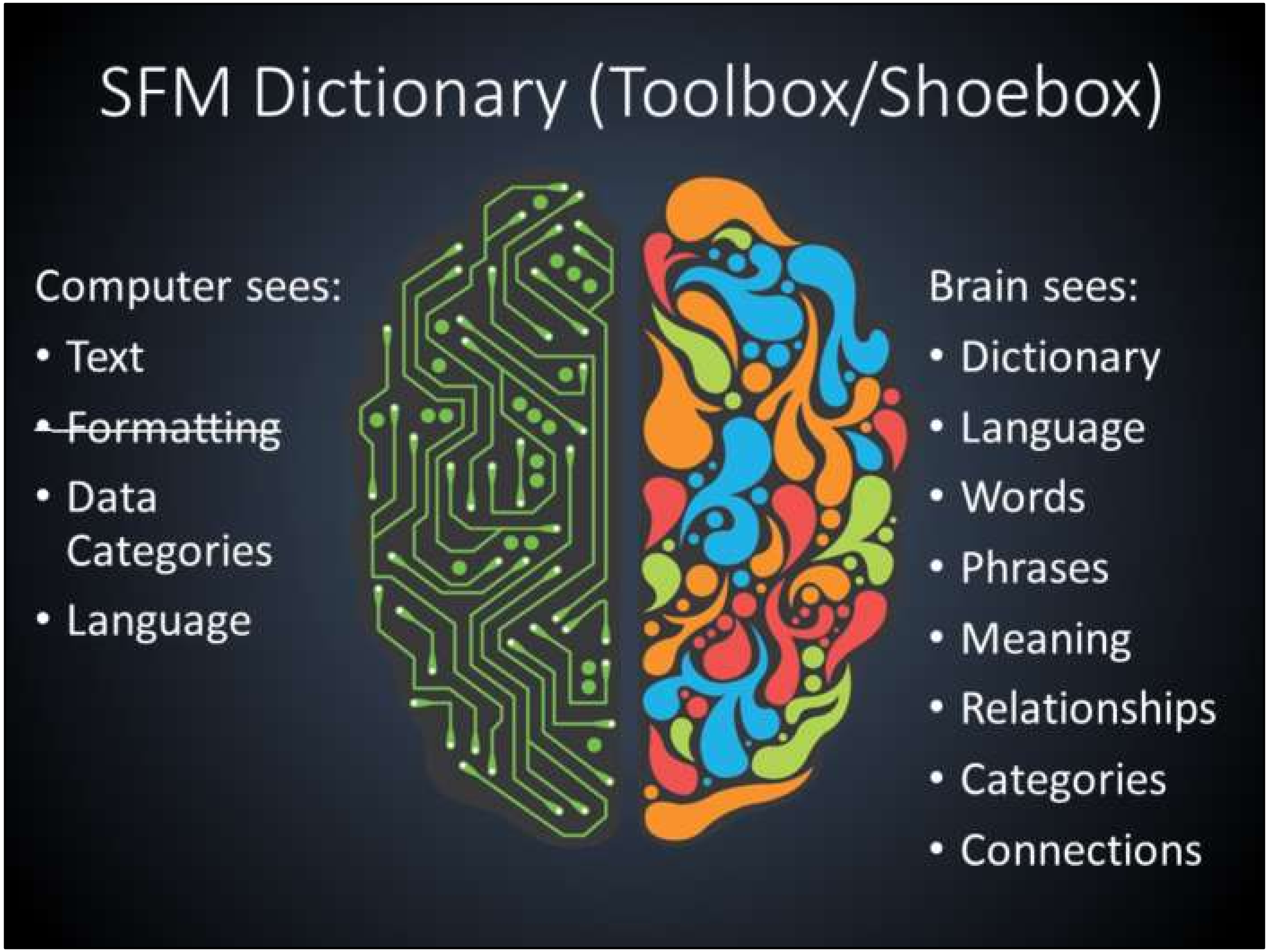
Image Source: http://www.tipsquirrel.com/photoshop-3d-whats-an-extrusion/

At best, publishing a toolbox Lexicon is like this:

Publishing a Toolbox Database

Image Source: http://www.industryweek.com/innovation/innovation-one-size-fits-one

and at it’s worst…



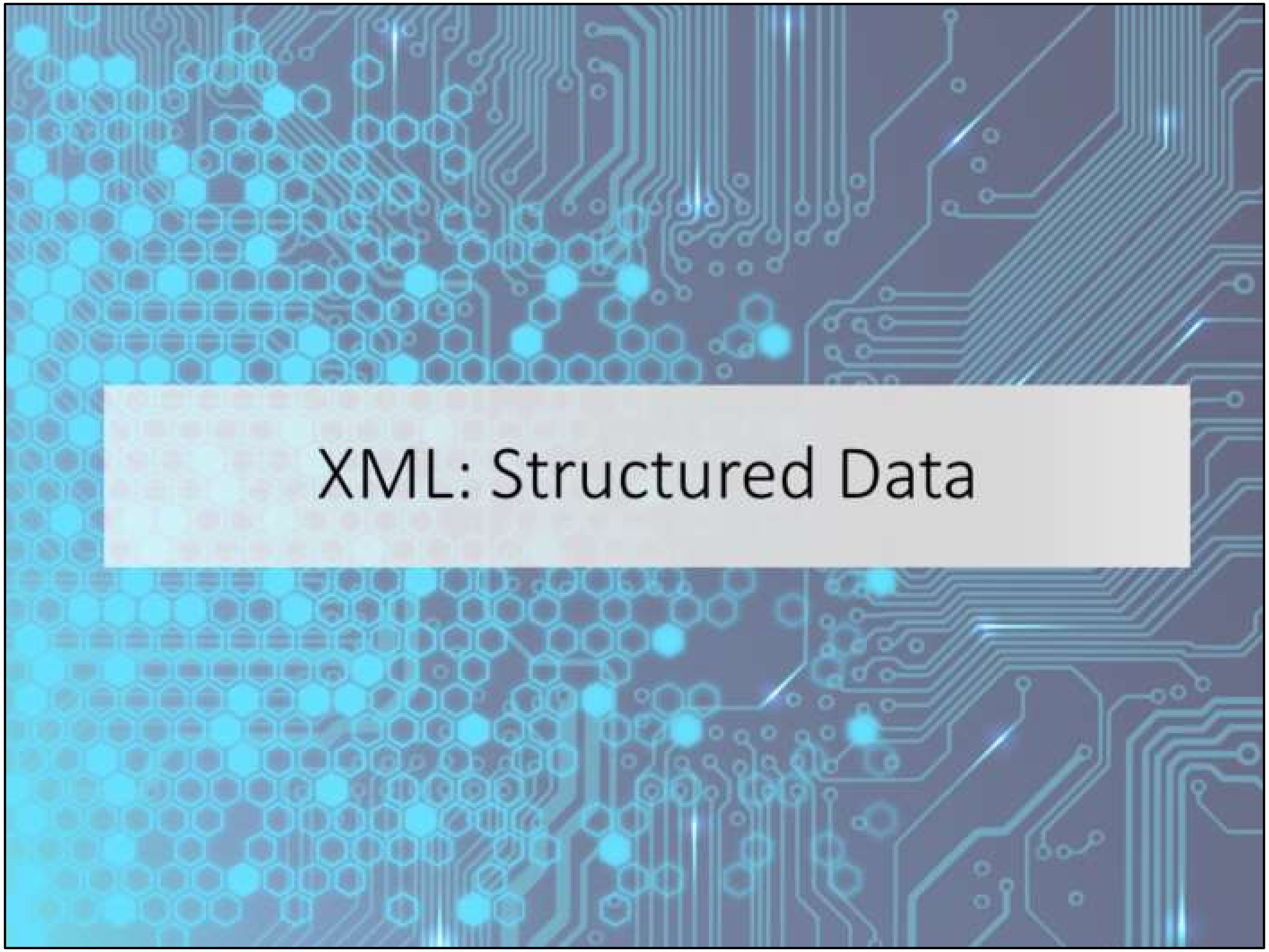
SFM Dictionary (Toolbox/Shoebox)

Computer sees:

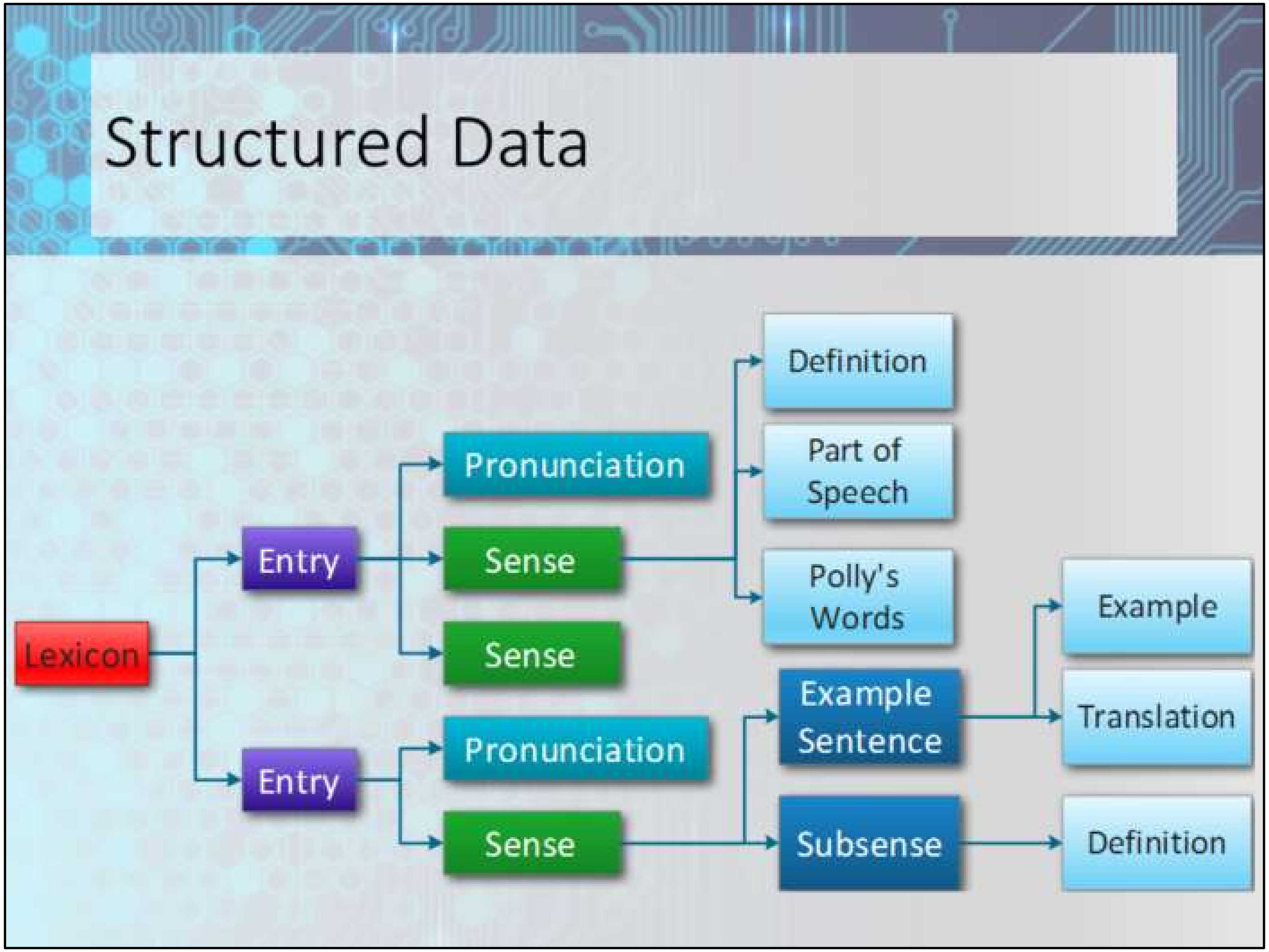
- Text
- ~~Formatting~~
- Data Categories
- Language

Brain sees:

- Dictionary
- Language
- Words
- Phrases
- Meaning
- Relationships
- Categories
- Connections



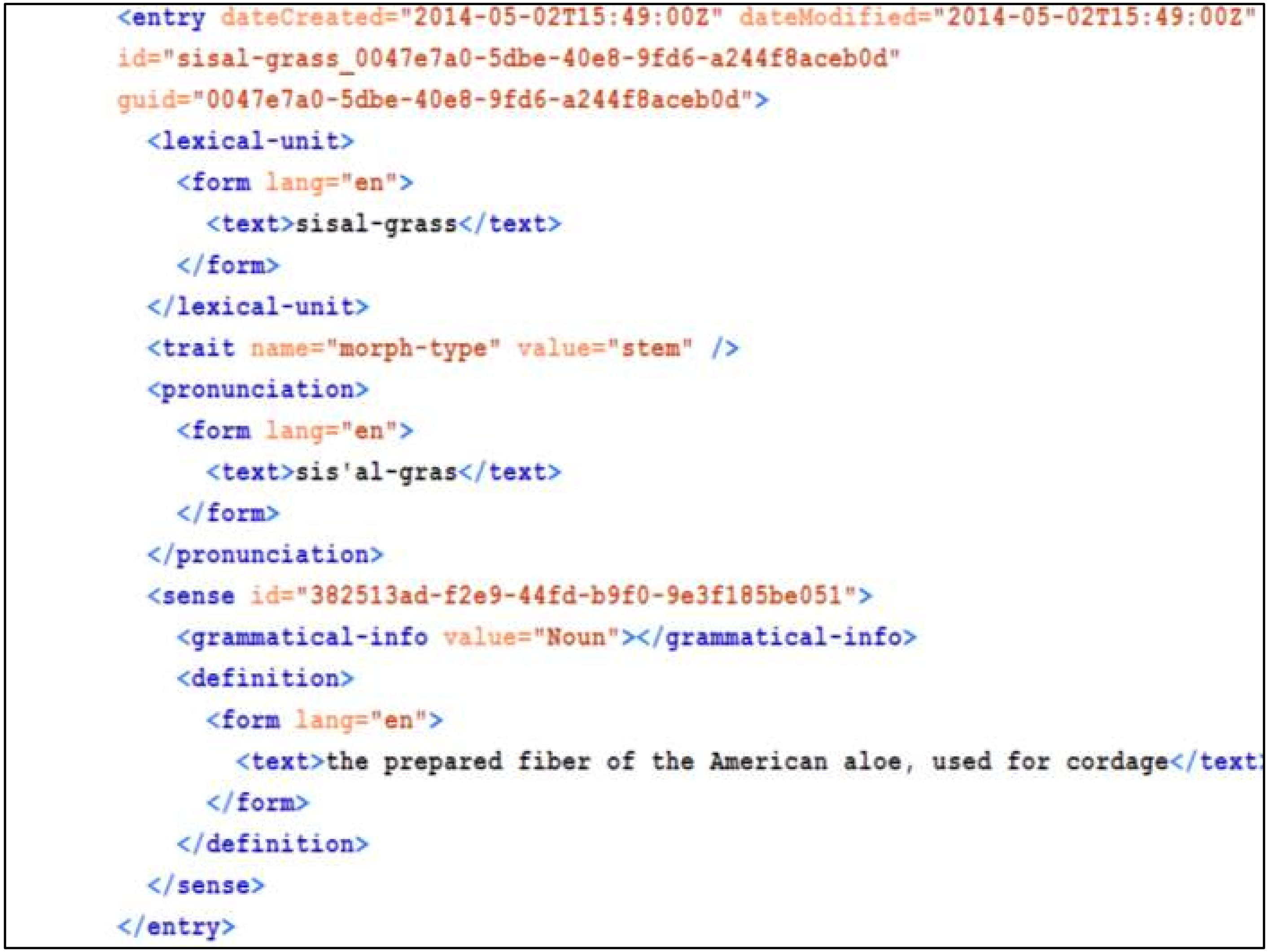
XML: Structured Data



A Lexicon has many Entries. Typically, each entry contains one or more senses and a pronunciation. Each sense can have zero or more example sentences. Each sense can have zero or more Subsenses. Even if you have a custom field…[click]…it has a place in the structure.

What does XML data look like?

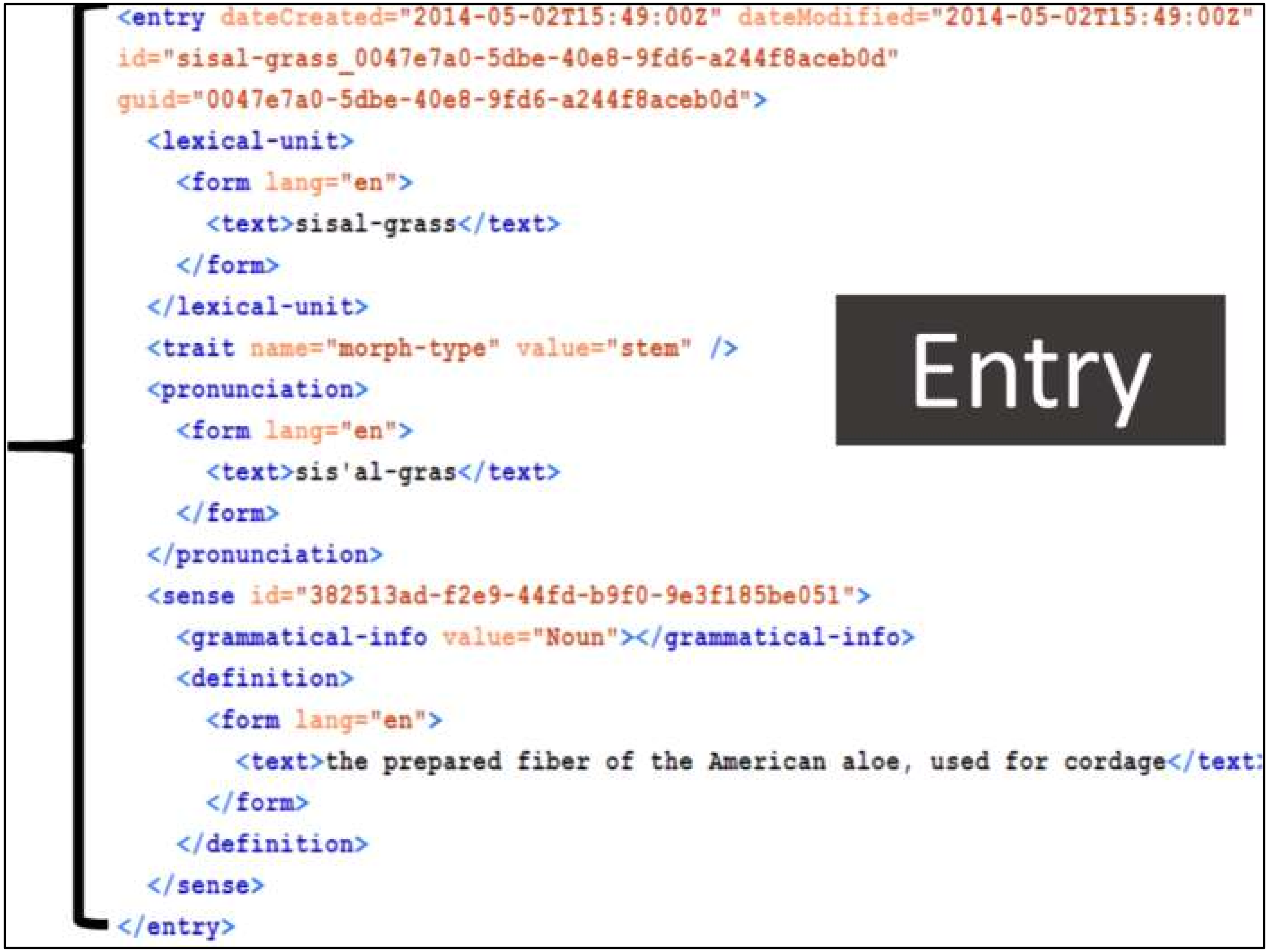
Dictionaries as Datapoints: Matthew Lee



```
<entry dateCreated="2014-05-02T15:49:00Z" dateModified="2014-05-02T15:49:00Z"
id="sisal-grass_0047e7a0-5dbe-40e8-9fd6-a244f8aceb0d"
guid="0047e7a0-5dbe-40e8-9fd6-a244f8aceb0d">
  <lexical-unit>
    <form lang="en">
      <text>sisal-grass</text>
    </form>
  </lexical-unit>
  <trait name="morph-type" value="stem" />
  <pronunciation>
    <form lang="en">
      <text>sis'al-gras</text>
    </form>
  </pronunciation>
  <sense id="382513ad-f2e9-44fd-b9f0-9e3f185be051">
    <grammatical-info value="Noun"></grammatical-info>
    <definition>
      <form lang="en">
        <text>the prepared fiber of the American aloe, used for cordage</text
      </form>
    </definition>
  </sense>
</entry>
```

LIFT XML looks like this…which may seem overwhelming at first, but it is self describing.

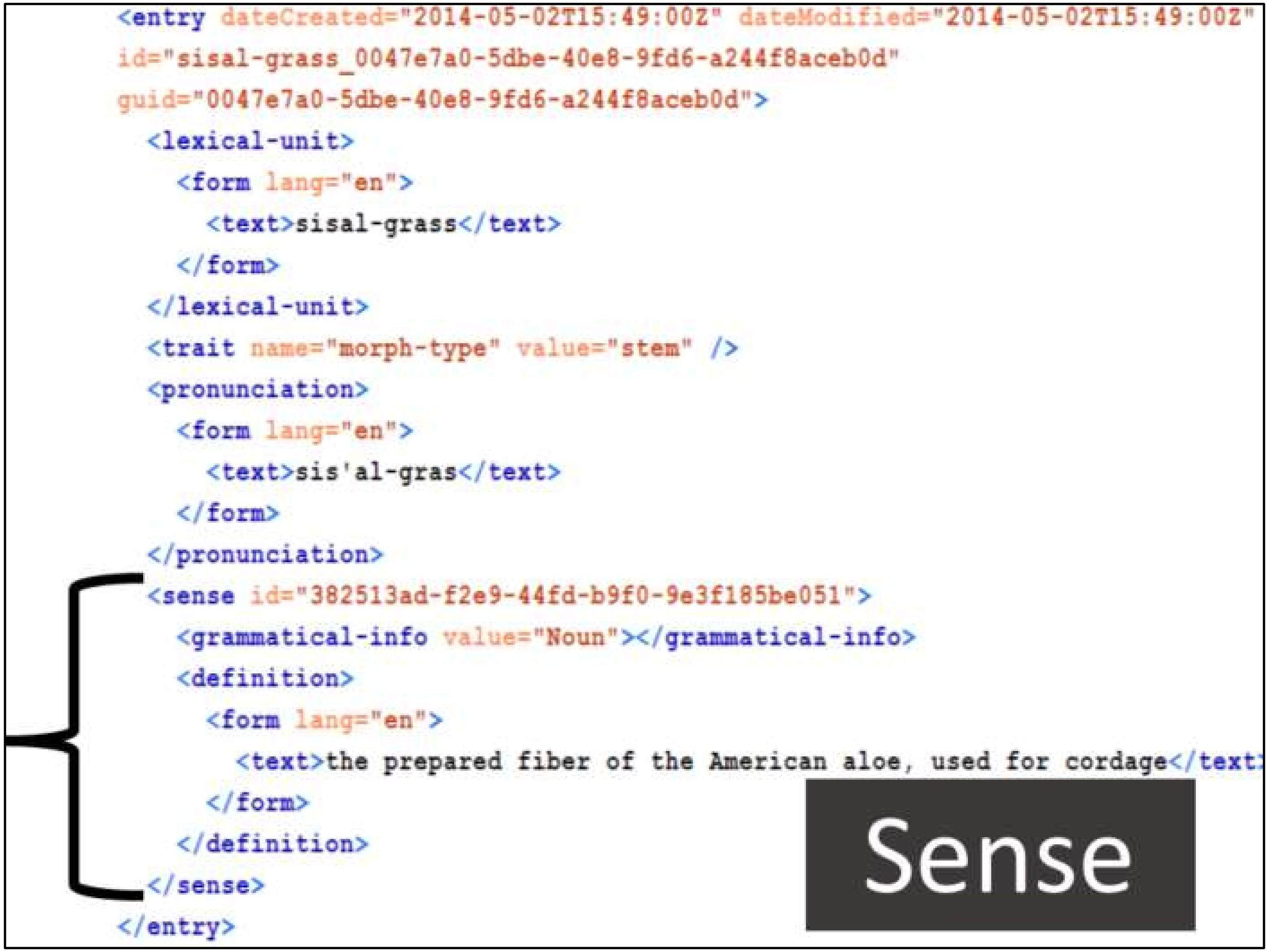
Dictionaries as Datapoints: Matthew Lee



```
<entry dateCreated="2014-05-02T15:49:00Z" dateModified="2014-05-02T15:49:00Z"
id="sisal-grass_0047e7a0-5dbe-40e8-9fd6-a244f8aceb0d"
guid="0047e7a0-5dbe-40e8-9fd6-a244f8aceb0d">
  <lexical-unit>
    <form lang="en">
      <text>sisal-grass</text>
    </form>
  </lexical-unit>
  <trait name="morph-type" value="stem" />
  <pronunciation>
    <form lang="en">
      <text>sis'al-gras</text>
    </form>
  </pronunciation>
  <sense id="382513ad-f2e9-44fd-b9f0-9e3f185be051">
    <grammatical-info value="Noun"></grammatical-info>
    <definition>
      <form lang="en">
        <text>the prepared fiber of the American aloe, used for cordage</text
      </form>
    </definition>
  </sense>
</entry>
```

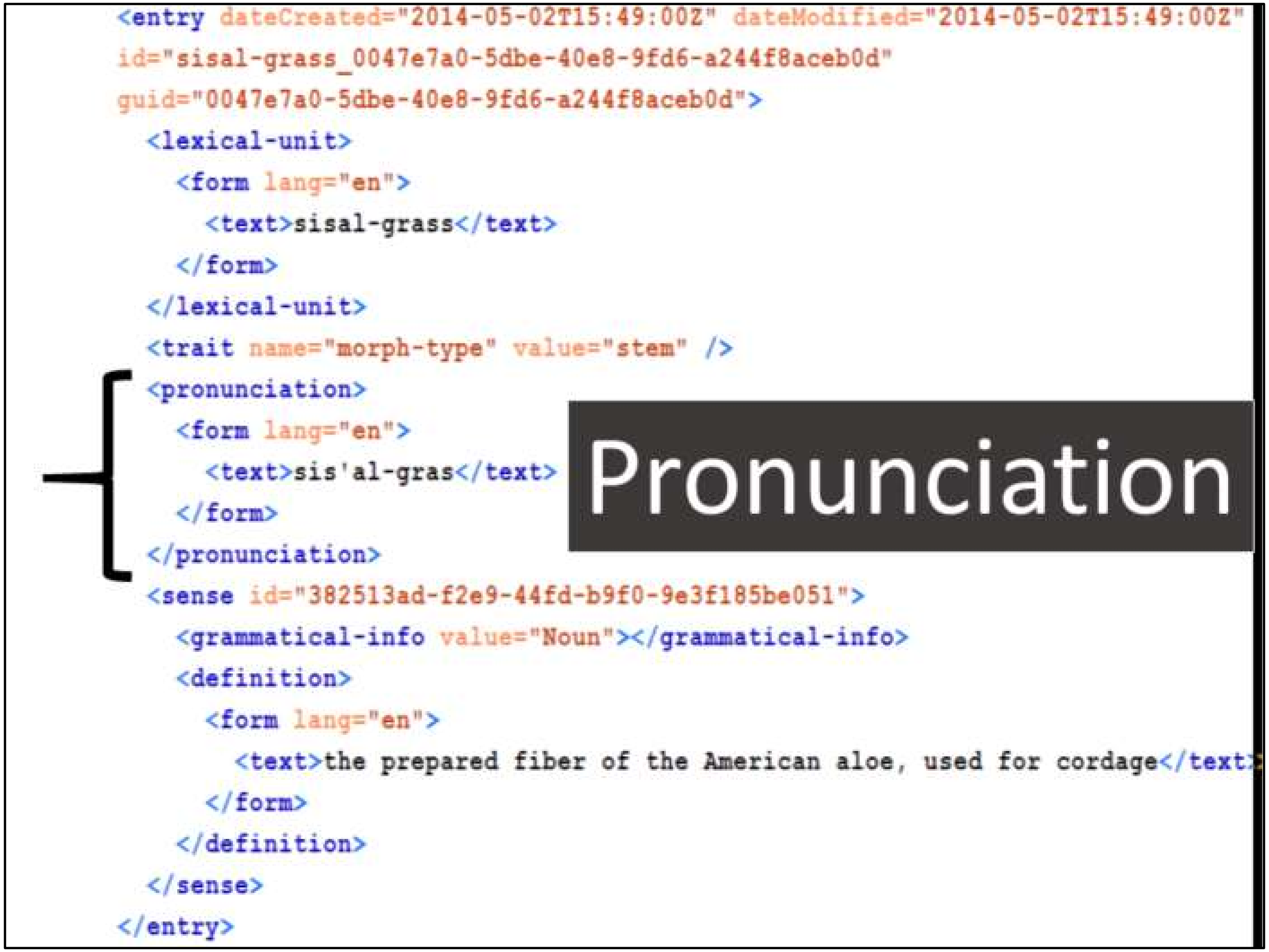
LIFT XML

Dictionaries as Datapoints: Matthew Lee



```
<entry dateCreated="2014-05-02T15:49:00Z" dateModified="2014-05-02T15:49:00Z"
id="sisal-grass_0047e7a0-5dbe-40e8-9fd6-a244f8aceb0d"
guid="0047e7a0-5dbe-40e8-9fd6-a244f8aceb0d">
  <lexical-unit>
    <form lang="en">
      <text>sisal-grass</text>
    </form>
  </lexical-unit>
  <trait name="morph-type" value="stem" />
  <pronunciation>
    <form lang="en">
      <text>sis'al-gras</text>
    </form>
  </pronunciation>
  <sense id="382513ad-f2e9-44fd-b9f0-9e3f185be051">
    <grammatical-info value="Noun"></grammatical-info>
    <definition>
      <form lang="en">
        <text>the prepared fiber of the American aloe, used for cordage</text
      </form>
    </definition>
  </sense>
</entry>
```

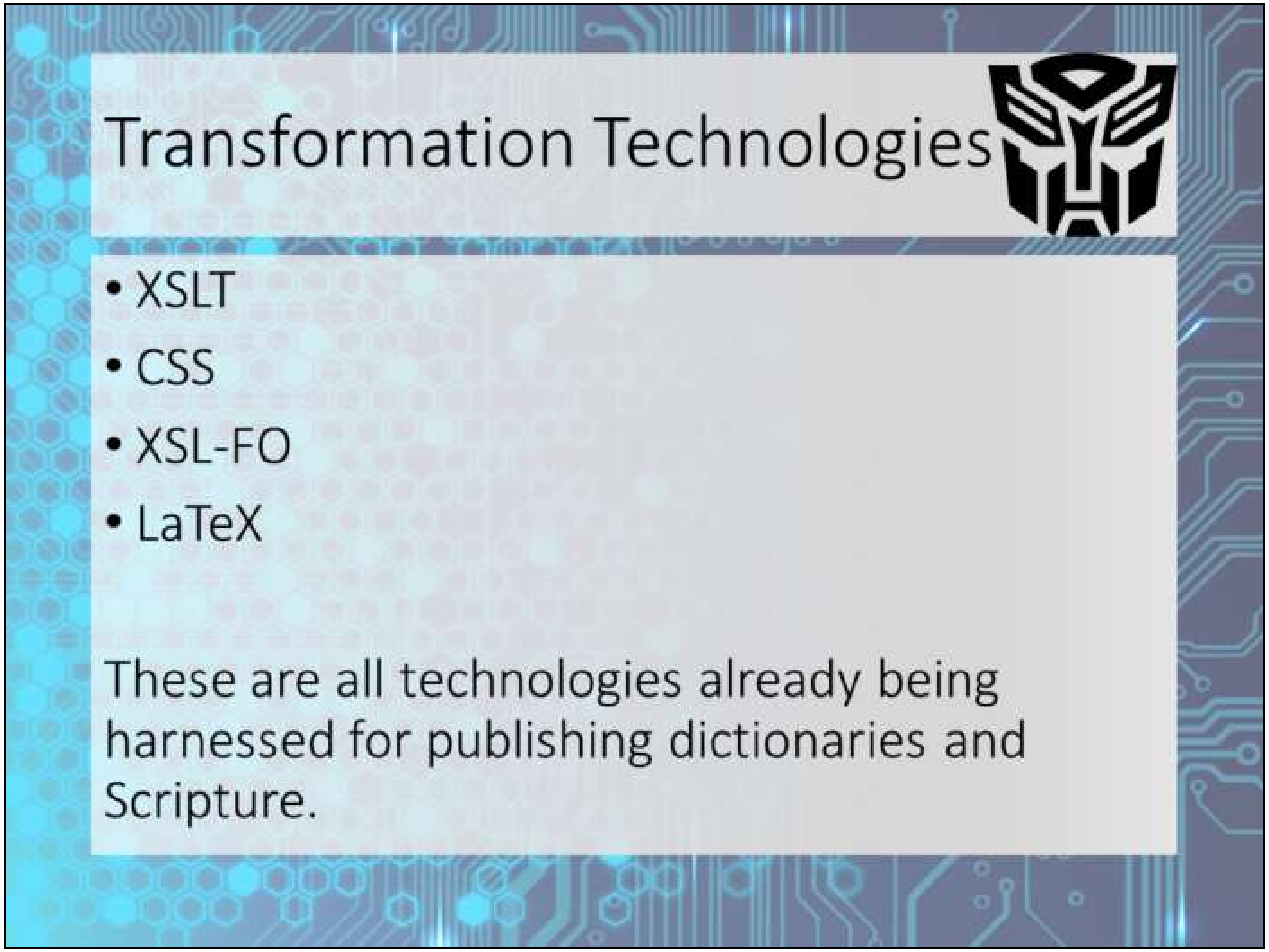
LIFT XML



```
<entry dateCreated="2014-05-02T15:49:00Z" dateModified="2014-05-02T15:49:00Z"
id="sisal-grass_0047e7a0-5dbe-40e8-9fd6-a244f8aceb0d"
guid="0047e7a0-5dbe-40e8-9fd6-a244f8aceb0d">
  <lexical-unit>
    <form lang="en">
      <text>sisal-grass</text>
    </form>
  </lexical-unit>
  <trait name="morph-type" value="stem" />
  <pronunciation>
    <form lang="en">
      <text>sis'al-gras</text>
    </form>
  </pronunciation>
  <sense id="382513ad-f2e9-44fd-b9f0-9e3f185be051">
    <grammatical-info value="Noun"></grammatical-info>
    <definition>
      <form lang="en">
        <text>the prepared fiber of the American aloe, used for cordage</text>
      </form>
    </definition>
  </sense>
</entry>
```

Pronunciation

LIFT XML



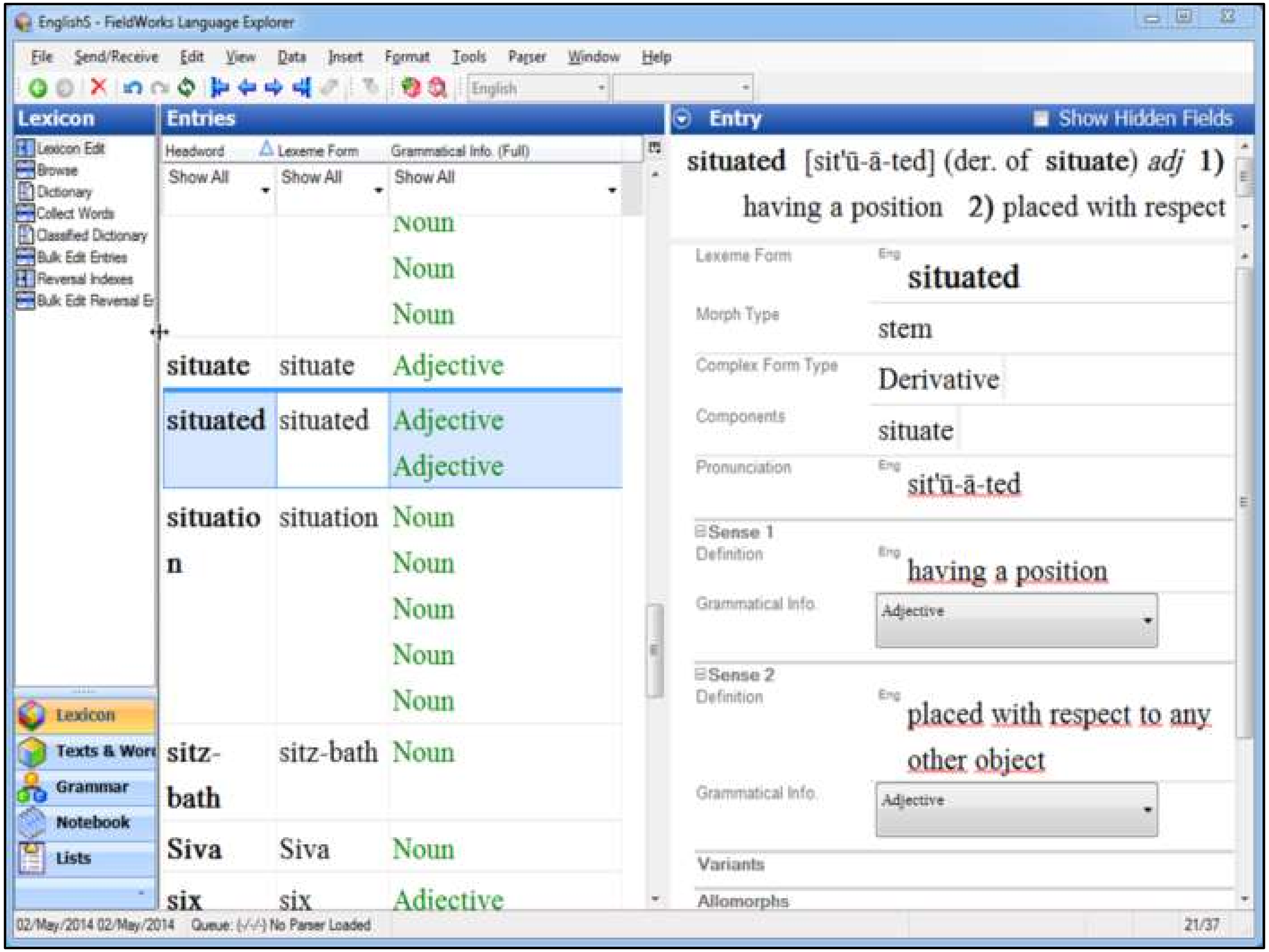
Transformation Technologies

- XSLT
- CSS
- XSL-FO
- LaTeX

These are all technologies already being harnessed for publishing dictionaries and Scripture.

Because the data is structured, it is relatively easy for the computer to reorganize the data and create a new form.

Dictionaries as Datapoints: Matthew Lee



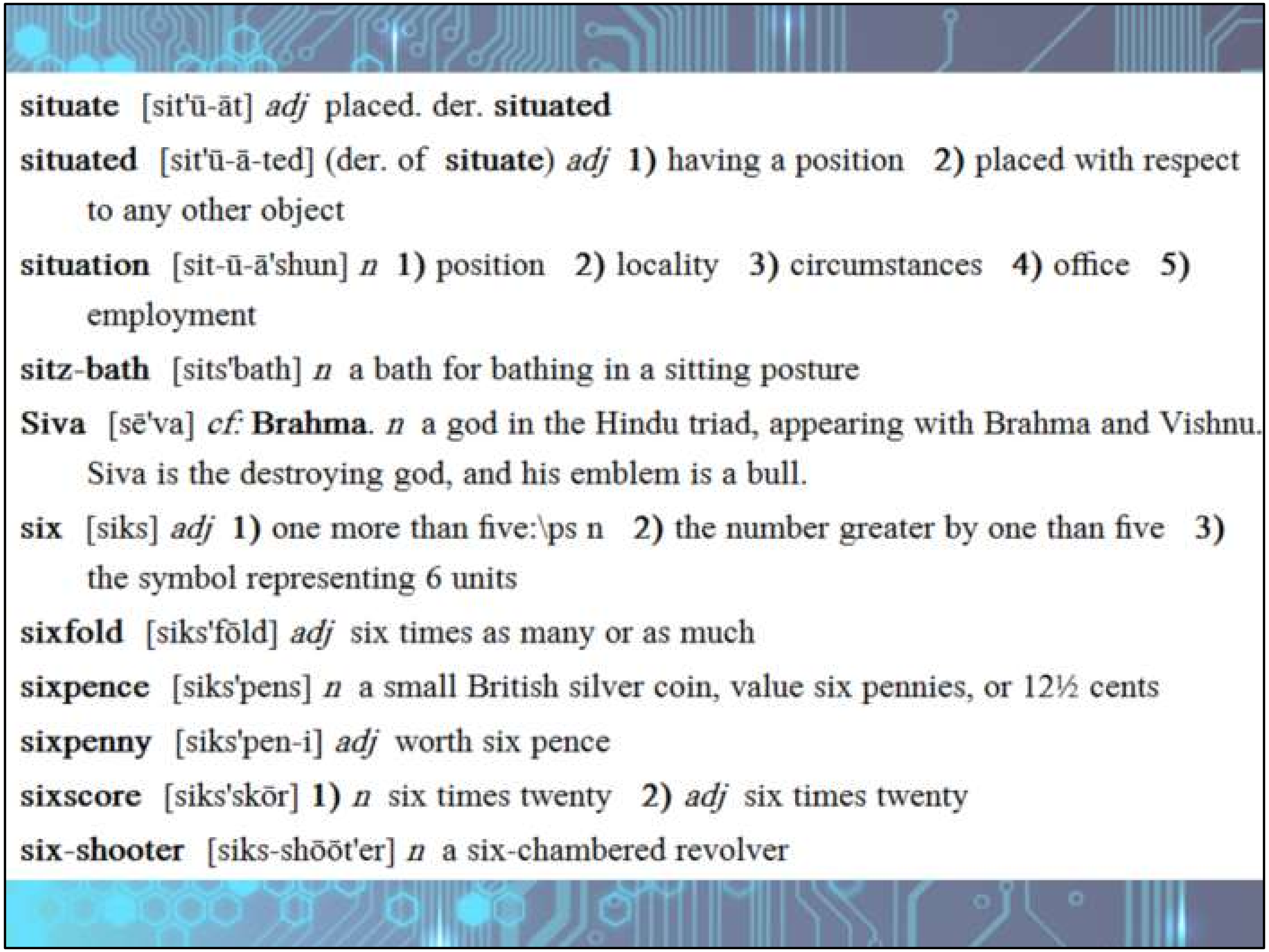
Headword	Lexeme Form	Grammatical Info. (Full)
Show All	Show All	Show All
		Noun Noun Noun
situate	situate	Adjective
situated	situated	Adjective Adjective
situation	situation	Noun Noun Noun Noun Noun
sitz-bath	sitz-bath	Noun
Siva	Siva	Noun
six	six	Adjective

situated [sit'ū-ā-ted] (der. of situate) *adj* 1) having a position 2) placed with respect

Lexeme Form	Eng situated
Morph Type	stem
Complex Form Type	Derivative
Components	situate
Pronunciation	Eng sit'ū-ā-ted
Sense 1 Definition	Eng having a position
Grammatical Info.	Adjective
Sense 2 Definition	Eng placed with respect to any other object
Grammatical Info.	Adjective
Variants	
Allomorphs	

02/May/2014 02/May/2014 Queue: (-/-/-) No Parser Loaded 21/37

This is SIL's Fieldworks Language Explorer. A new generation of linguists has been trained on FLEx, and many are transitioning from Toolbox to FLEx. FLEx uses XML to store it's data and offers fields for most of the categories that you will need, with the option of adding custom ones into the structure. **XML Data can look like this…[click]**



situate [sit'ū-āt] *adj* placed. der. **situated**

situated [sit'ū-ā-ted] (der. of **situate**) *adj* **1)** having a position **2)** placed with respect to any other object

situation [sit-ū-ā'shun] *n* **1)** position **2)** locality **3)** circumstances **4)** office **5)** employment

sitz-bath [sits'bath] *n* a bath for bathing in a sitting posture

Siva [sē'va] *cf.* **Brahma**. *n* a god in the Hindu triad, appearing with Brahma and Vishnu. Siva is the destroying god, and his emblem is a bull.

six [siks] *adj* **1)** one more than five:\ps n **2)** the number greater by one than five **3)** the symbol representing 6 units

sixfold [siks'fōld] *adj* six times as many or as much

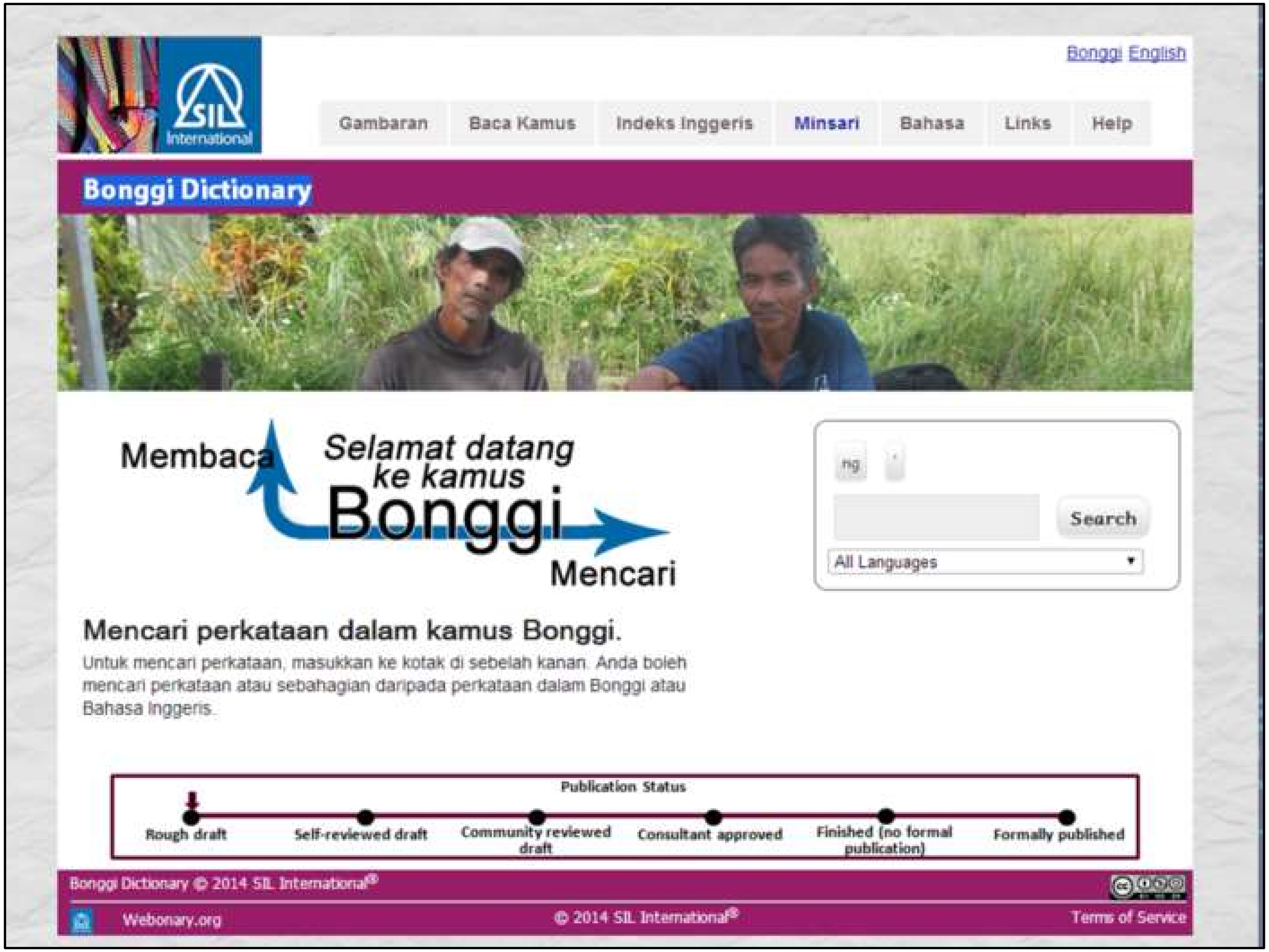
sixpence [siks'pens] *n* a small British silver coin, value six pennies, or 12½ cents

sixpenny [siks'pen-i] *adj* worth six pence

sixscore [siks'skōr] **1)** *n* six times twenty **2)** *adj* six times twenty

six-shooter [siks-shōōt'er] *n* a six-chambered revolver

Or this



Mencari perkataan dalam kamus Bonggi.

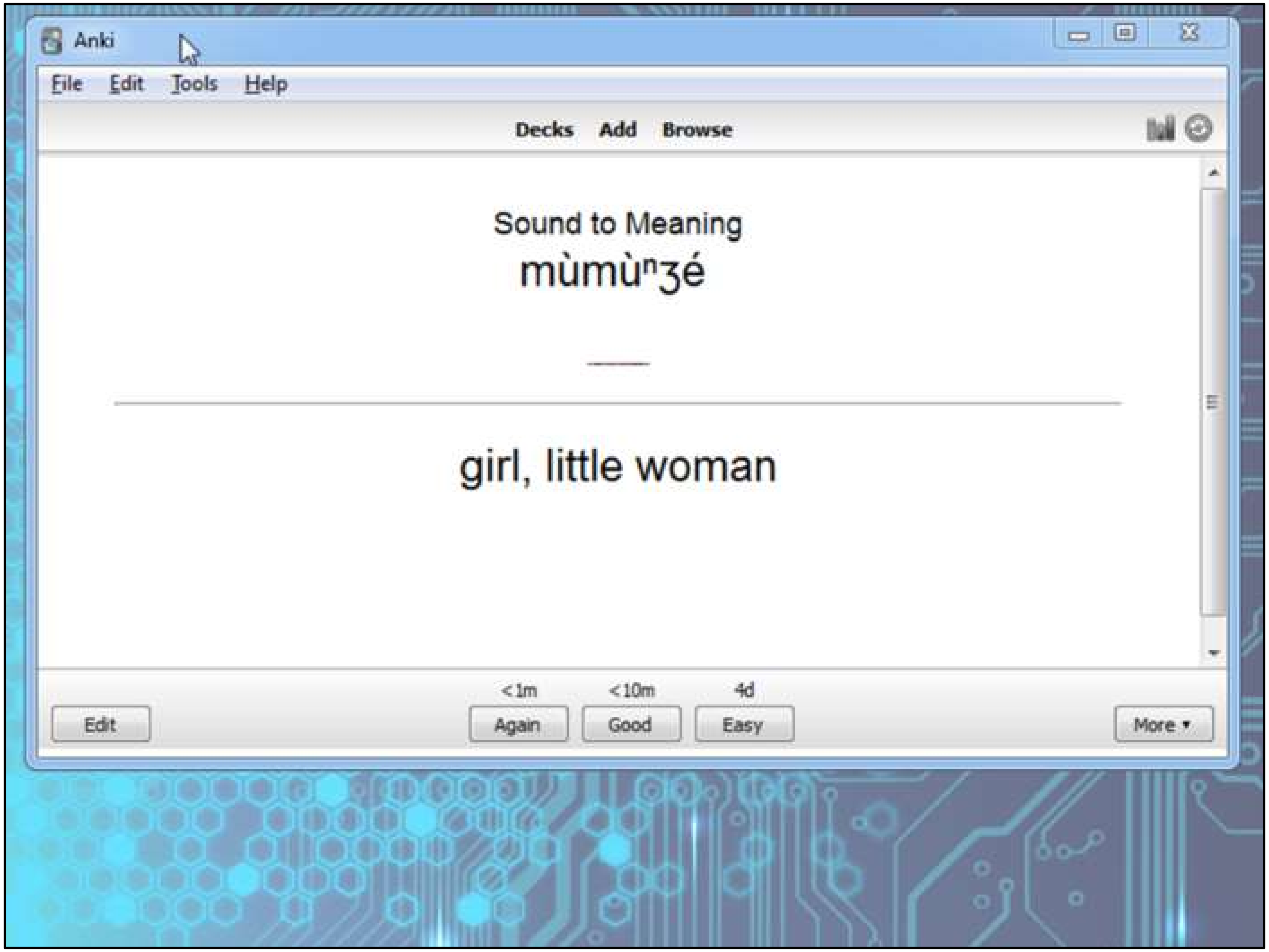
Untuk mencari perkataan, masukkan ke kotak di sebelah kanan. Anda boleh mencari perkataan atau sebahagian daripada perkataan dalam Bonggi atau Bahasa Inggeris.

Bonggi Dictionary © 2014 SIL International®

© 2014 SIL International®

Or this (Bonngi Webonary)

Dictionaries as Datapoints: Matthew Lee



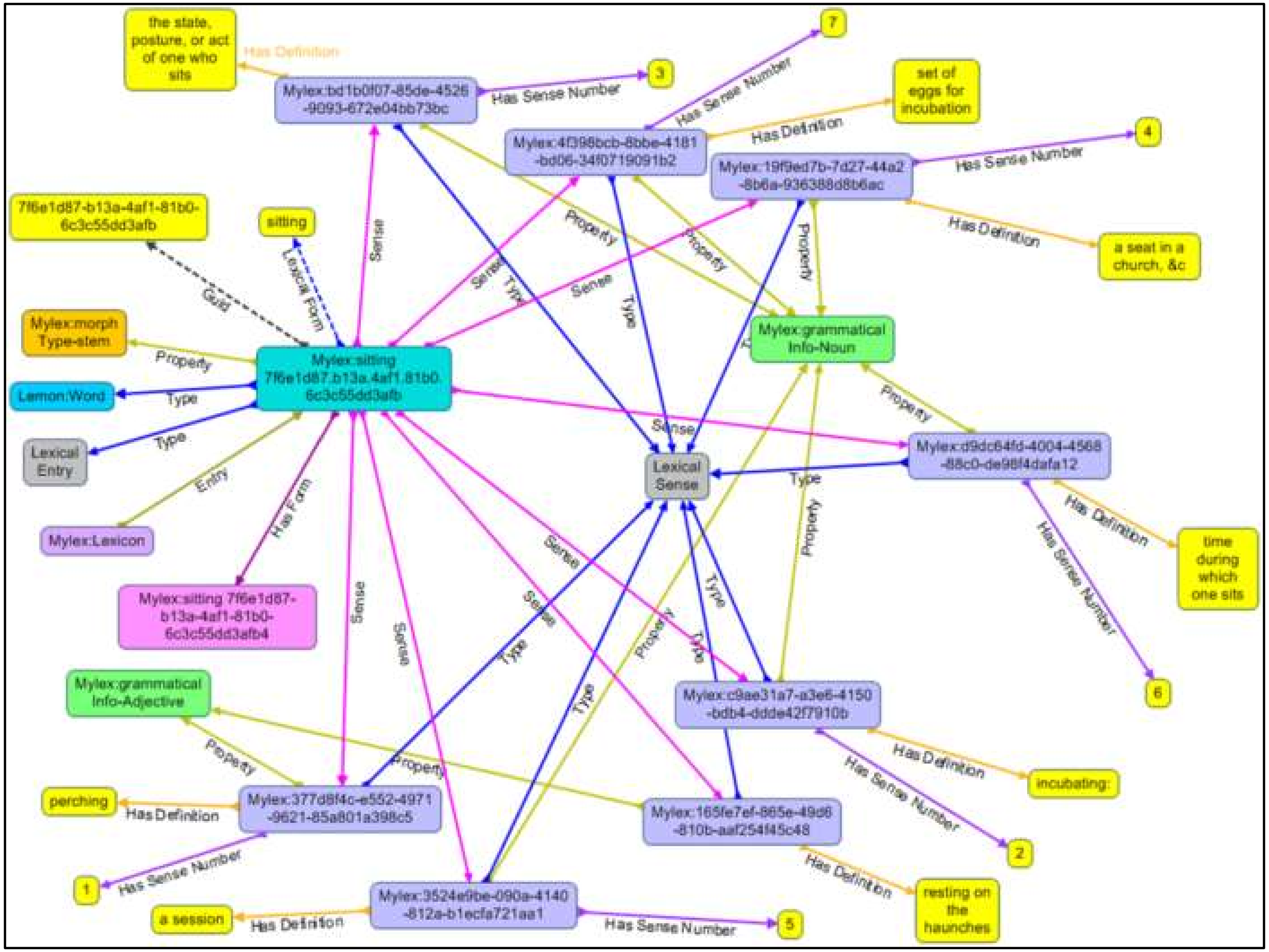
Sound to Meaning

mùmùⁿʒé

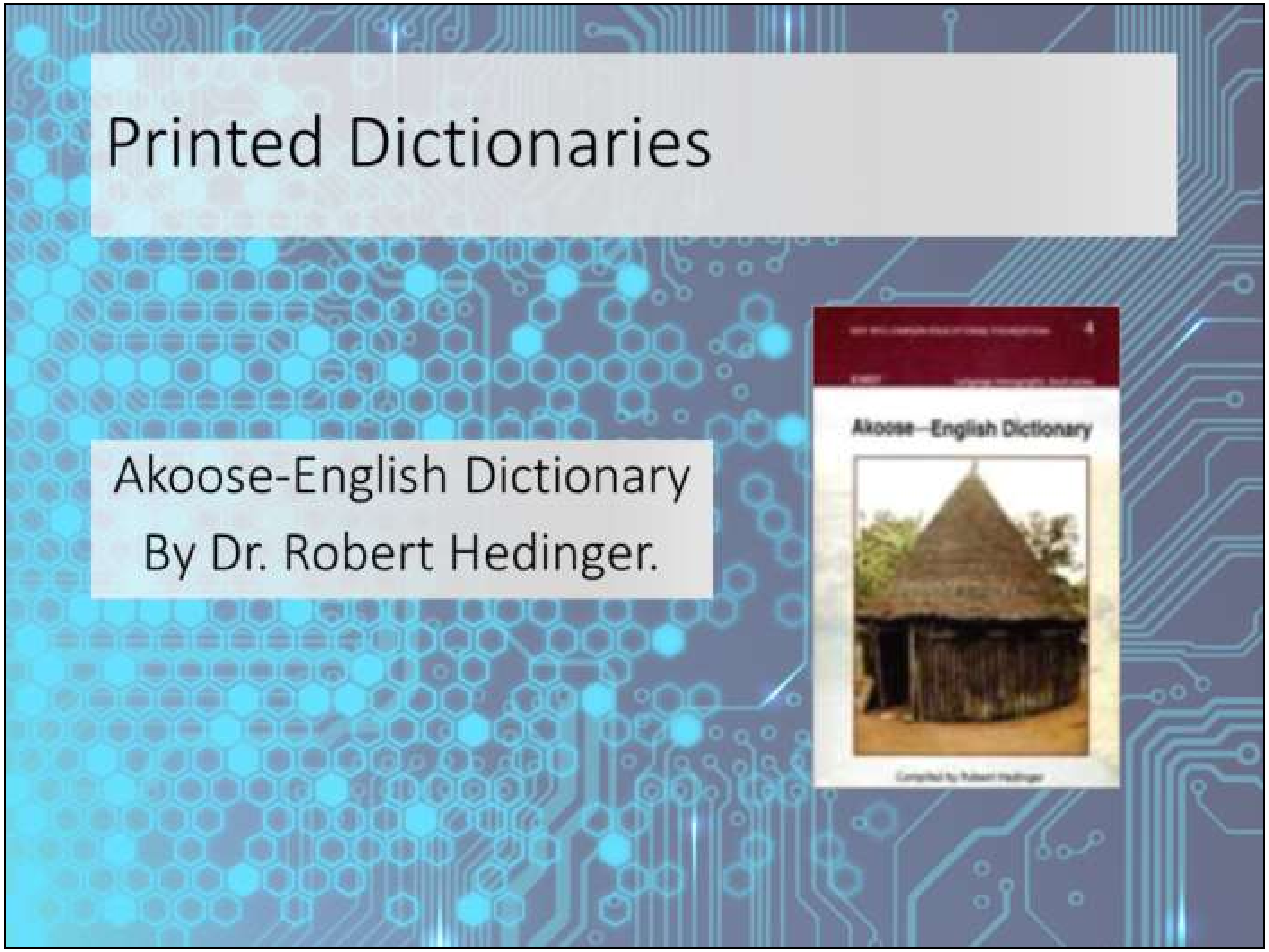
girl, little woman

Or this (Anki Flashcards)

Dictionaries as Datapoints: Matthew Lee


the state, posture, or act of one who sits
Has Definition
Mylex:bd1b0f07-85de-4526-9093-672e04bb73bc
Has Sense Number
3
7
Has Sense Number
set of eggs for incubation
Has Definition
Mylex:4f398bcb-8bbe-4181-bd06-34f0719091b2
Mylex:19f9ed7b-7d27-44a2-8b6a-936388d8b6ac
Has Sense Number
4
7f6e1d87-b13a-4af1-81b0-6c3c55dd3afb
sitting
Has Definition
a seat in a church, &c
Guid
Lexical Form
Sense
Sense
Type
Property
Sense
Type
Property
Property
Mylex:morph Type-stem
Property
Mylex:sitting 7f6e1d87.b13a.4af1.81b0.6c3c55dd3afb
Mylex:grammatical Info-Noun
Lemon:Word
Type
Type
Lexical Entry
Entry
Sense
Property
Mylex:d9dc64fd-4004-4568-88c0-de98f4dafa12
Lexical Sense
Type
Mylex:Lexicon
Has Form
Property
Has Definition
Has Sense Number
time during which one sits
Mylex:sitting 7f6e1d87-b13a-4af1-81b0-6c3c55dd3afb4
Sense
Sense
Sense
Sense
Type
Type
Property
Type
Mylex:grammatical Info-Adjective
Mylex:c9ae31a7-a3e6-4150-bdb4-ddde42f7910b
6
Property
Property
Type
Has Definition
incubating;
perching
Has Definition
Mylex:377d8f4c-e552-4971-9621-85a801a398c5
Mylex:165fe7ef-865e-49d6-810b-aaf254f45c48
Has Sense Number
2
1
Has Sense Number
Has Definition
resting on the haunches
Mylex:3524e9be-090a-4140-812a-b1ecfa721aa1
a session
Has Definition
Has Sense Number
5

I’ll come back to this one..

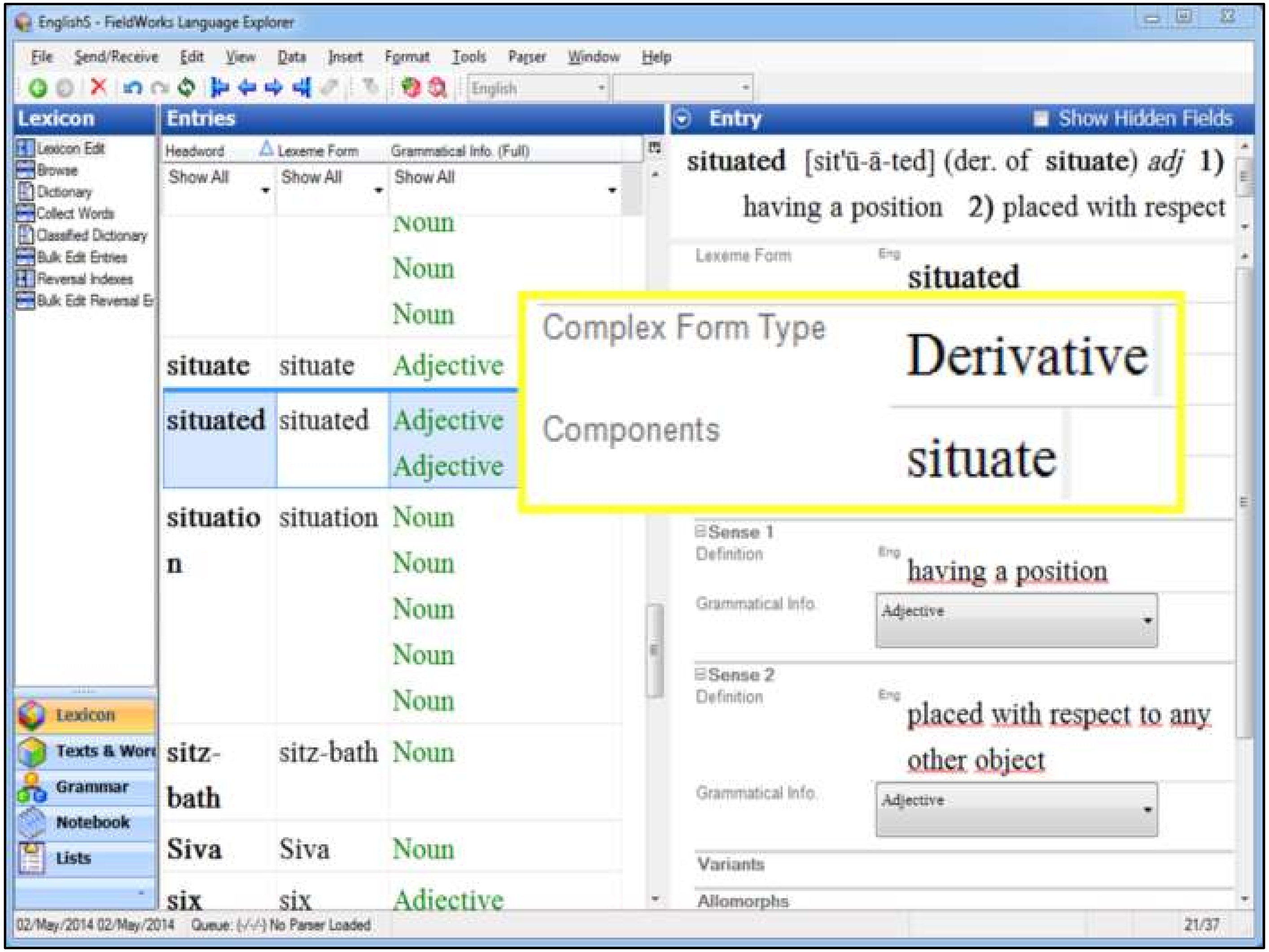


Printed Dictionaries

Akoose-English Dictionary
By Dr. Robert Hedinger.

Or of course it can be printed…

Dictionaries as Datapoints: Matthew Lee



Headword	Lexeme Form	Grammatical Info. (Full)
Show All	Show All	Show All
		Noun Noun Noun
situate	situate	Adjective
situated	situated	Adjective Adjective
situatio n	situation	Noun Noun Noun Noun Noun
sitz- bath	sitz-bath	Noun
Siva	Siva	Noun
six	six	Adjective

situated [sit'ū-ā-ted] (der. of **situate**) *adj* **1)** having a position **2)** placed with respect

Lexeme Form Eng situated

Complex Form Type Derivative

Components situate

Sense 1

Definition Eng having a position

Grammatical Info. Adjective

Sense 2

Definition Eng placed with respect to any other object

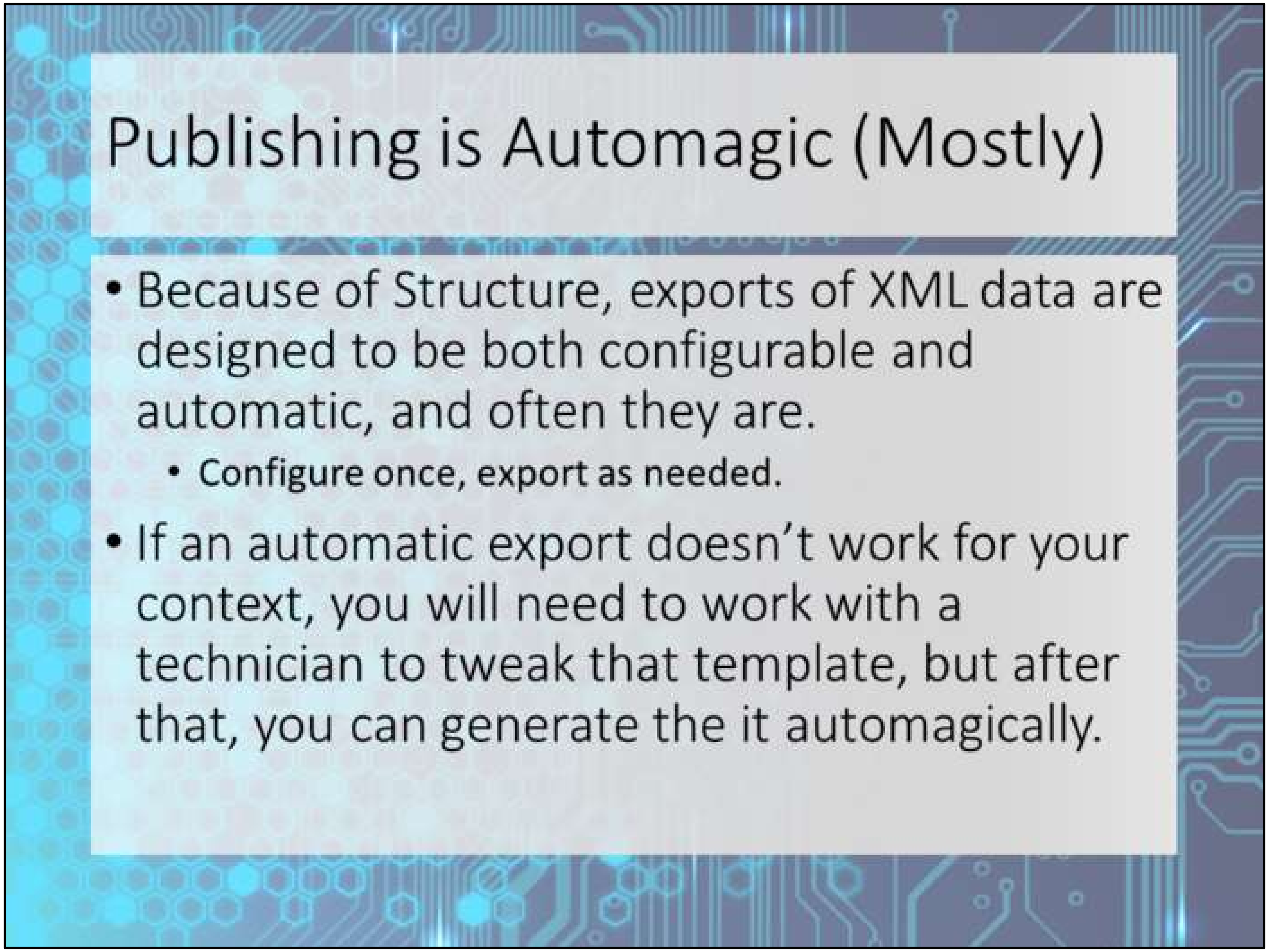
Grammatical Info. Adjective

Variants

Allomorphs

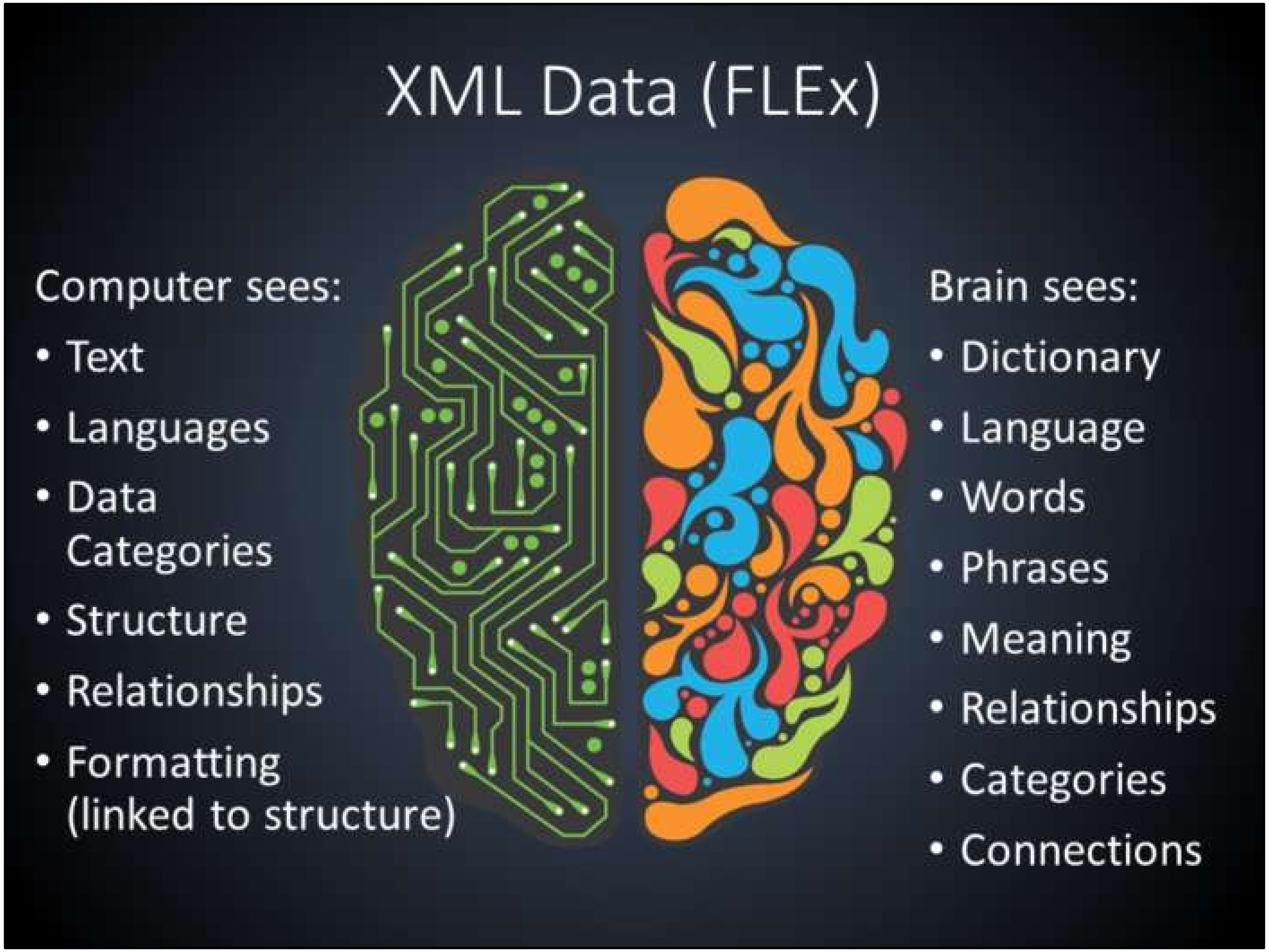
02/May/2014 02/May/2014 Queue: (-/-/-) No Parser Loaded 21/37

XML data can also be linked…this shows that the adjective situate is related to “situate”. The tool makes sure that the target entry exists to prevent dead links.



Publishing is Automagic (Mostly)

- Because of Structure, exports of XML data are designed to be both configurable and automatic, and often they are.
  - Configure once, export as needed.
- If an automatic export doesn’t work for your context, you will need to work with a technician to tweak that template, but after that, you can generate the it automagically.



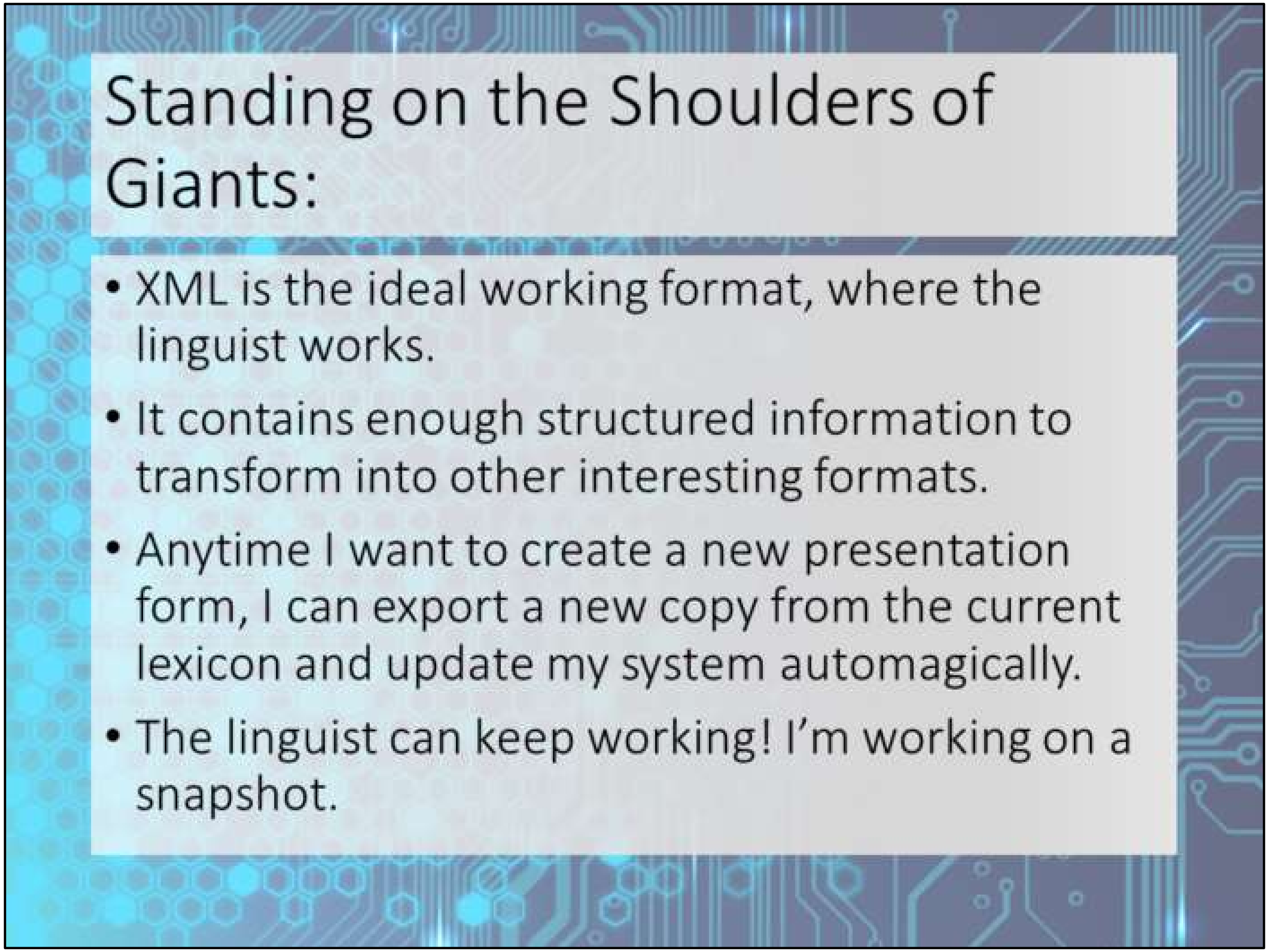
XML Data (FLEx)

Computer sees:

- Text
- Languages
- Data Categories
- Structure
- Relationships
- Formatting (linked to structure)

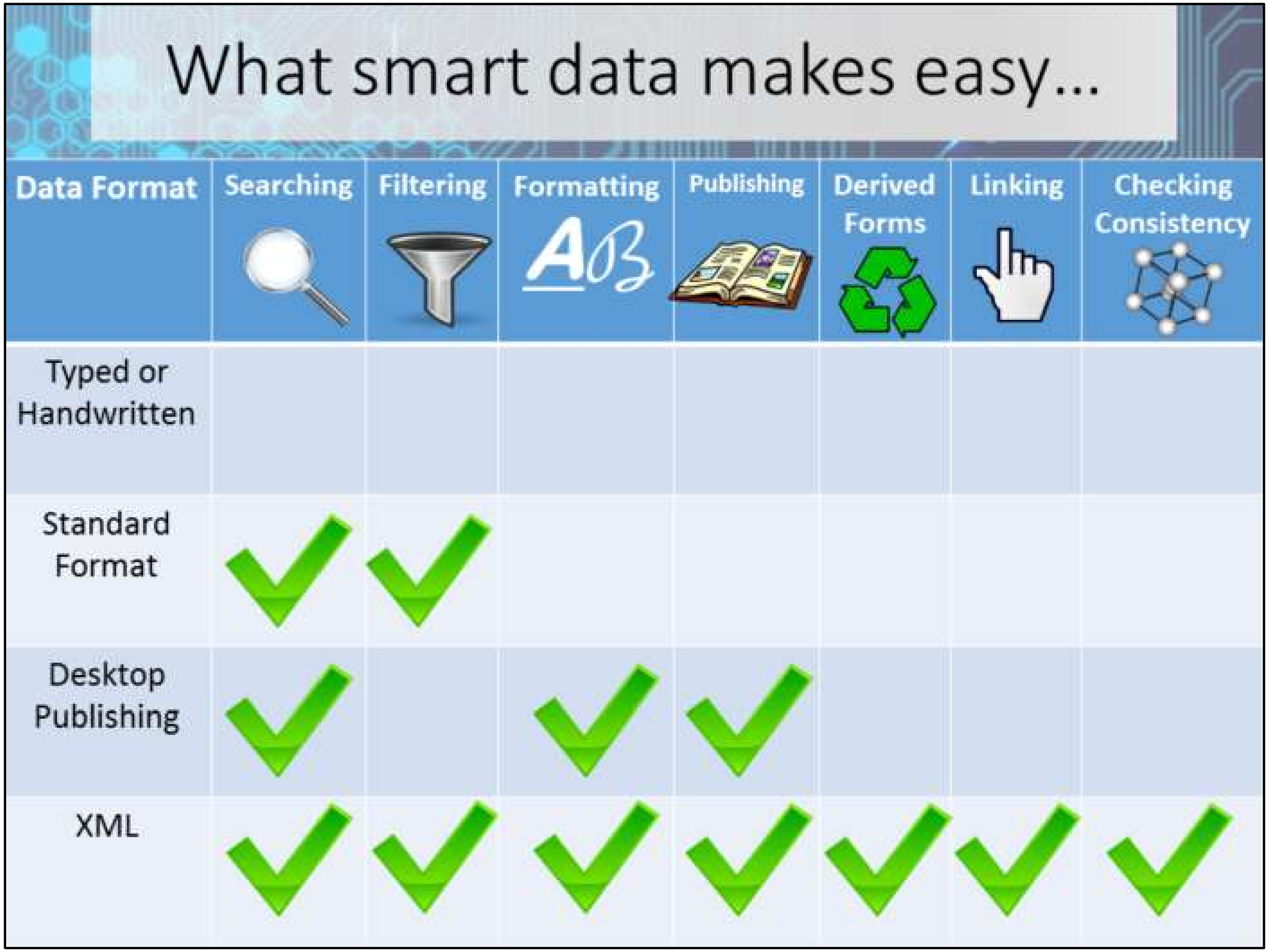
Brain sees:

- Dictionary
- Language
- Words
- Phrases
- Meaning
- Relationships
- Categories
- Connections



Standing on the Shoulders of Giants:

- XML is the ideal working format, where the linguist works.
- It contains enough structured information to transform into other interesting formats.
- Anytime I want to create a new presentation form, I can export a new copy from the current lexicon and update my system automagically.
- The linguist can keep working! I'm working on a snapshot.



What smart data makes easy…

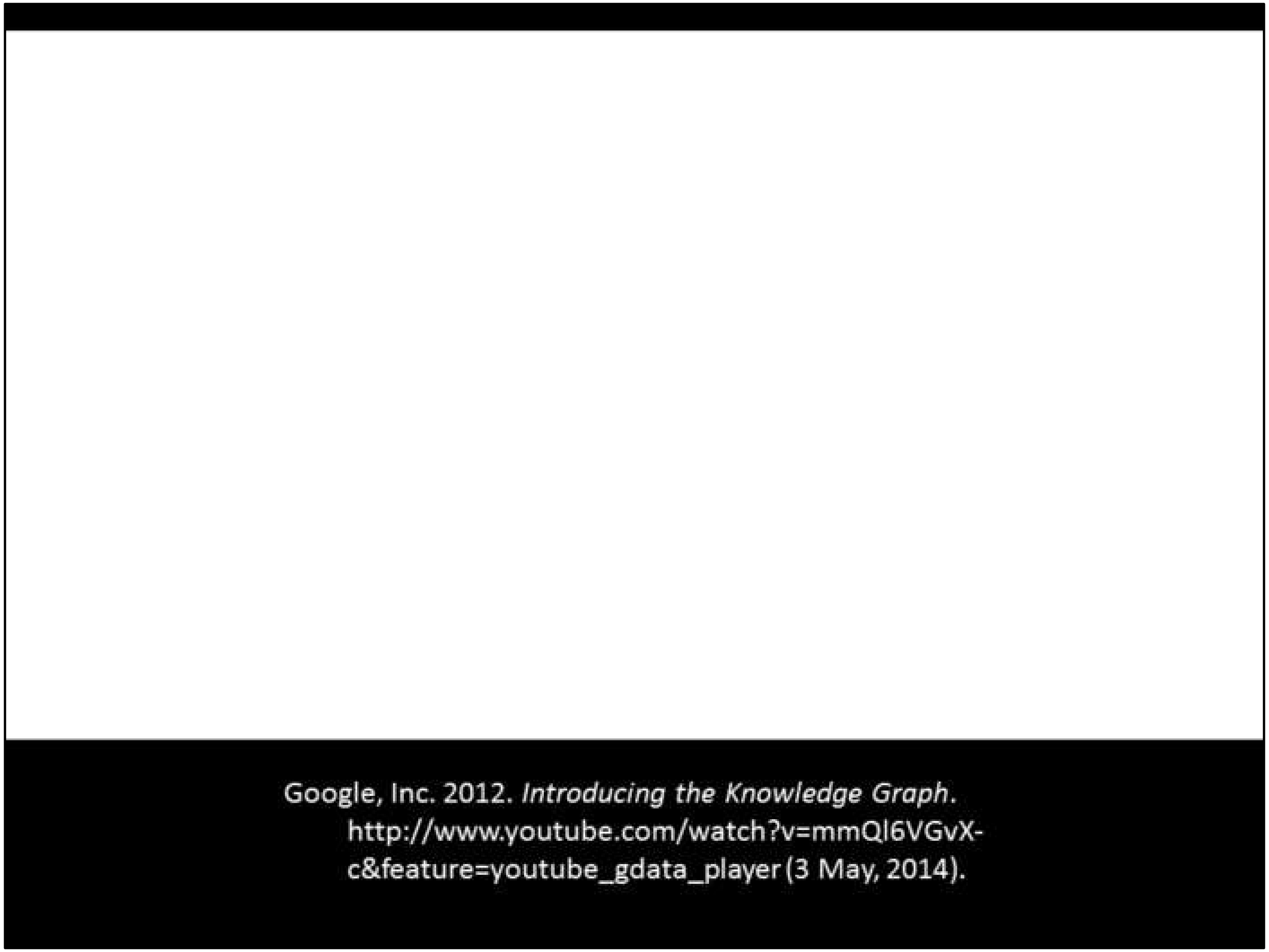
Data Format	Searching	Filtering	Formatting	Publishing	Derived Forms	Linking	Checking Consistency
Typed or Handwritten							
Standard Format	✓	✓					
Desktop Publishing	✓		✓	✓			
XML	✓	✓	✓	✓	✓	✓	✓

This chart shows what these types of data are good at, though it could be argued that each of these things is possible. If I had room, I'd add "Leveling your Coffee Table" so that the handwritten dictionary could have a point in its favor.

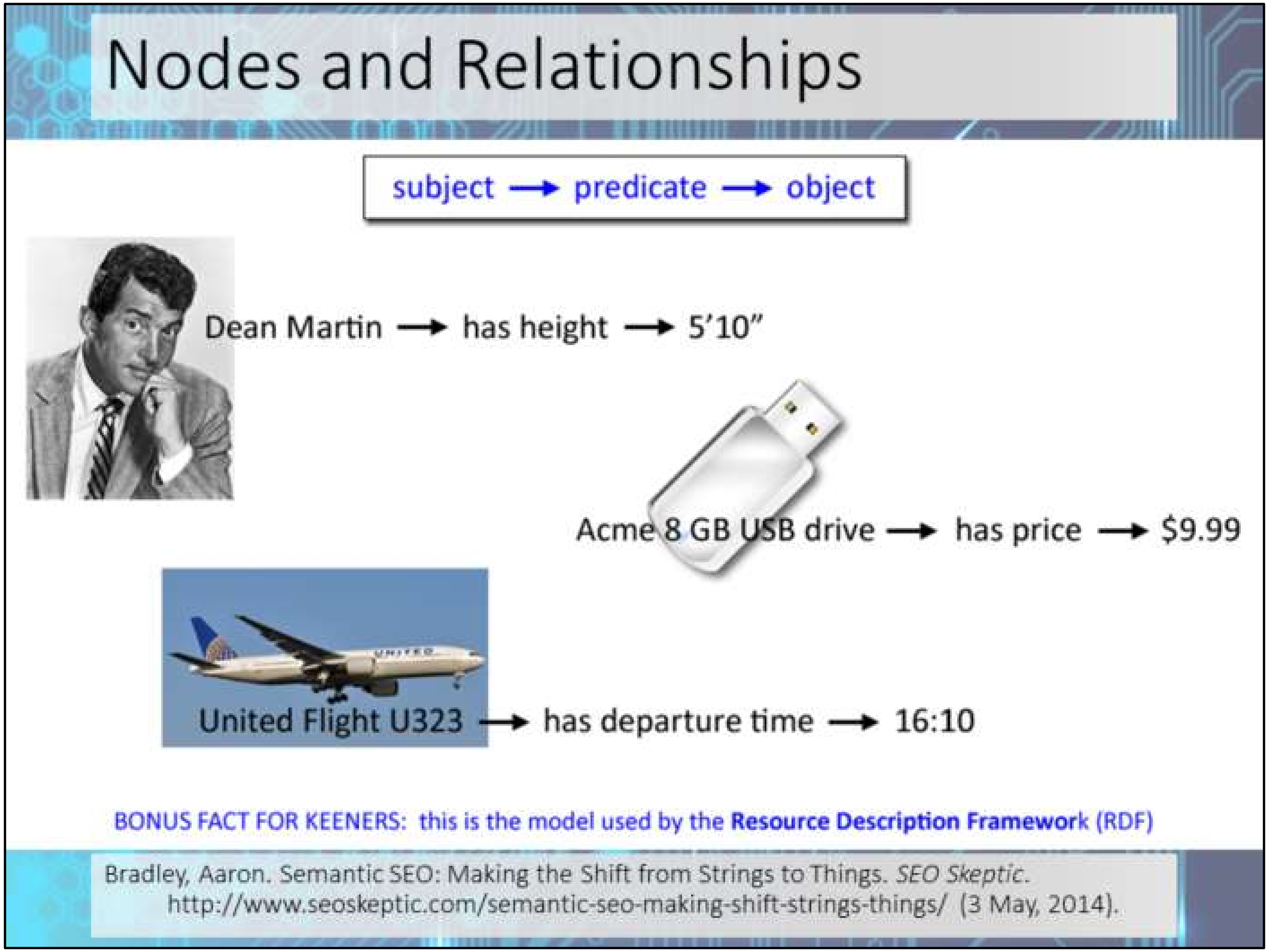
RDF: Resource Description Format

The Technology that I used to link the databases is called Resource Description Format. Here’s a video about a little company that is exploring this technology.

Dictionaries as Datapoints: Matthew Lee



Google, Inc. 2012. *Introducing the Knowledge Graph*.
http://www.youtube.com/watch?v=mmQl6VGvX-c&feature=youtube_gdata_player (3 May, 2014).



Nodes and Relationships

subject → predicate → object

Dean Martin → has height → 5’10”

Acme 8 GB USB drive → has price → $9.99

United Flight U323 → has departure time → 16:10

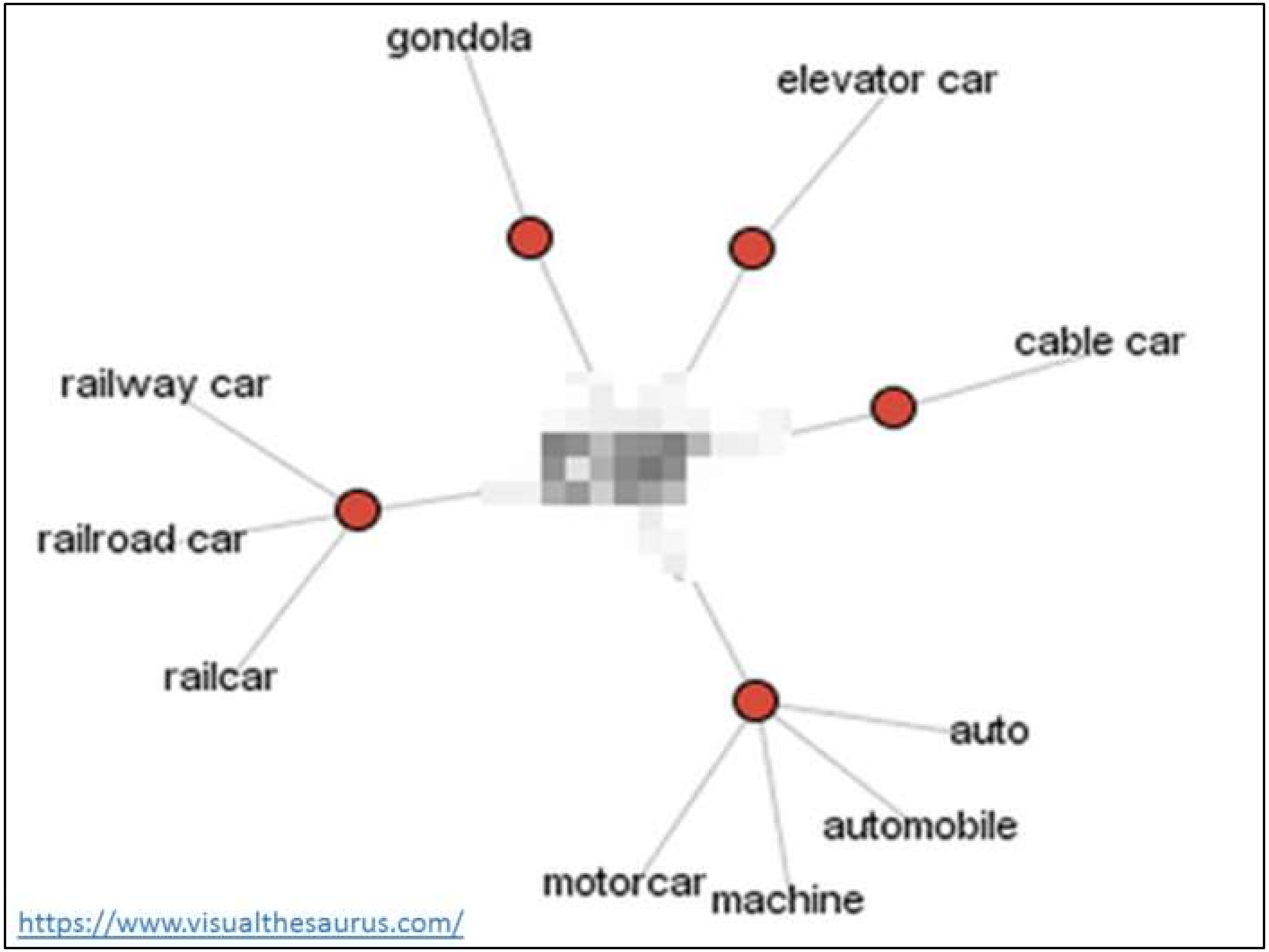
BONUS FACT FOR KEENERS: this is the model used by the **Resource Description Framework** (RDF)

Bradley, Aaron. Semantic SEO: Making the Shift from Strings to Things. *SEO Skeptic*. http://www.seoskeptic.com/semantic-seo-making-shift-strings-things/ (3 May, 2014).

All Knowledge can be represented by nodes and relationships.

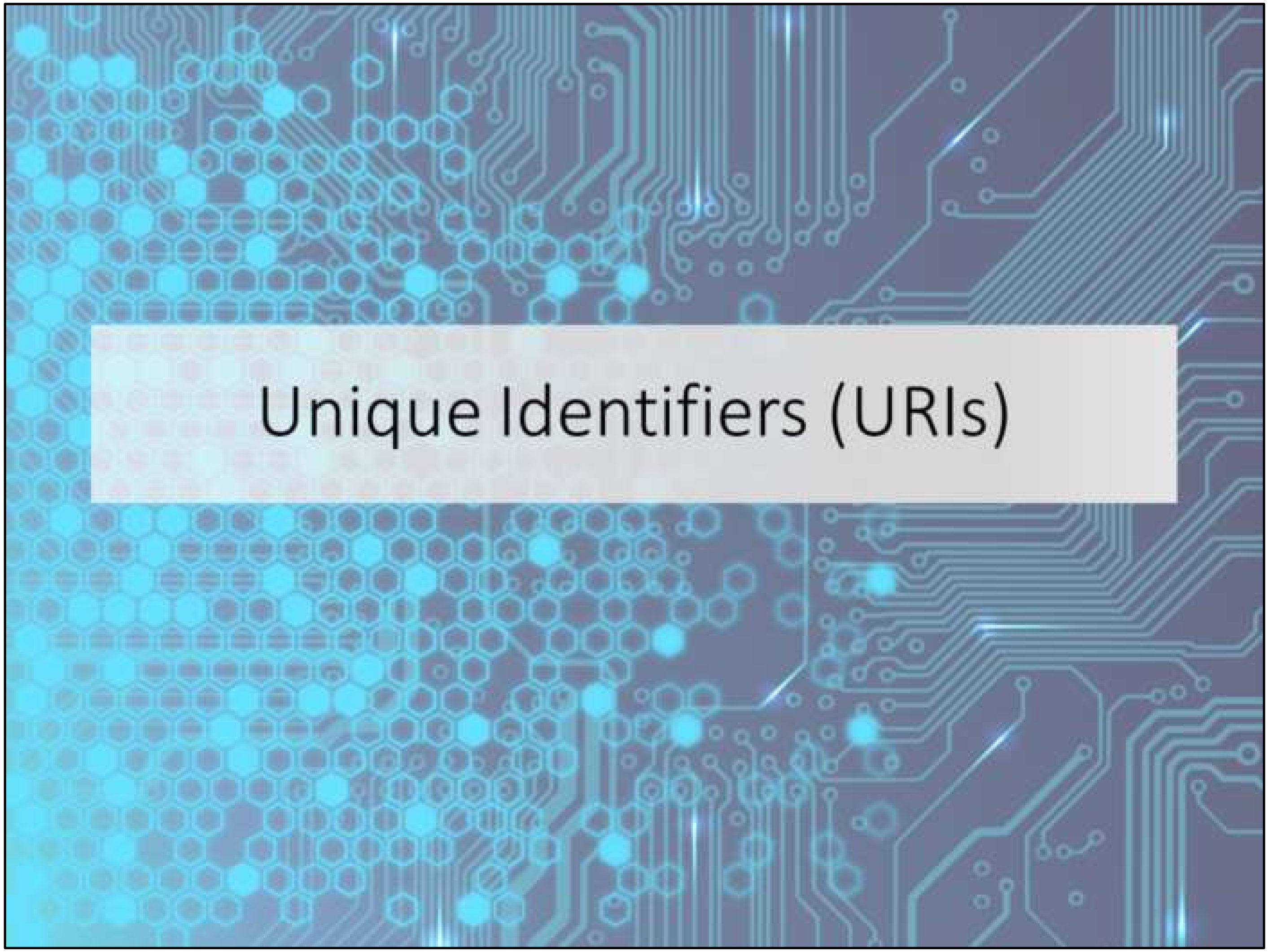
The more relationships…the more you know about a concept. If you don’t believe that everything can be defined by relationships…

Dictionaries as Datapoints: Matthew Lee



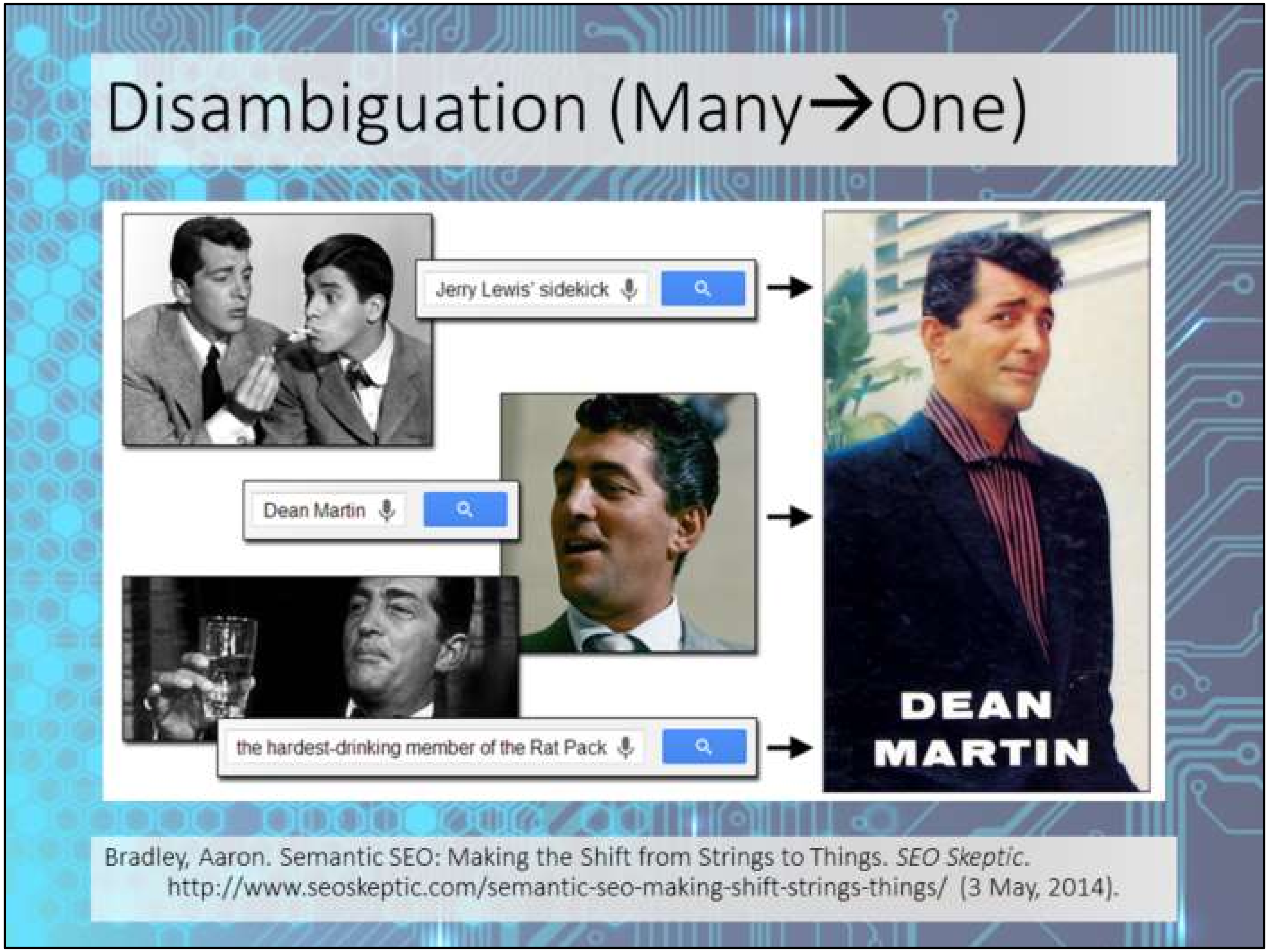
https://www.visualthesaurus.com/

If you don’t believe that everything can be defined by relationships… Take a Thesaurus… Each line represents “synonym”. Even if you don’t know the word you could learn something about it by studying the related words. The computer can help us understand things by showing those relationships.



Unique Identifiers (URIs)

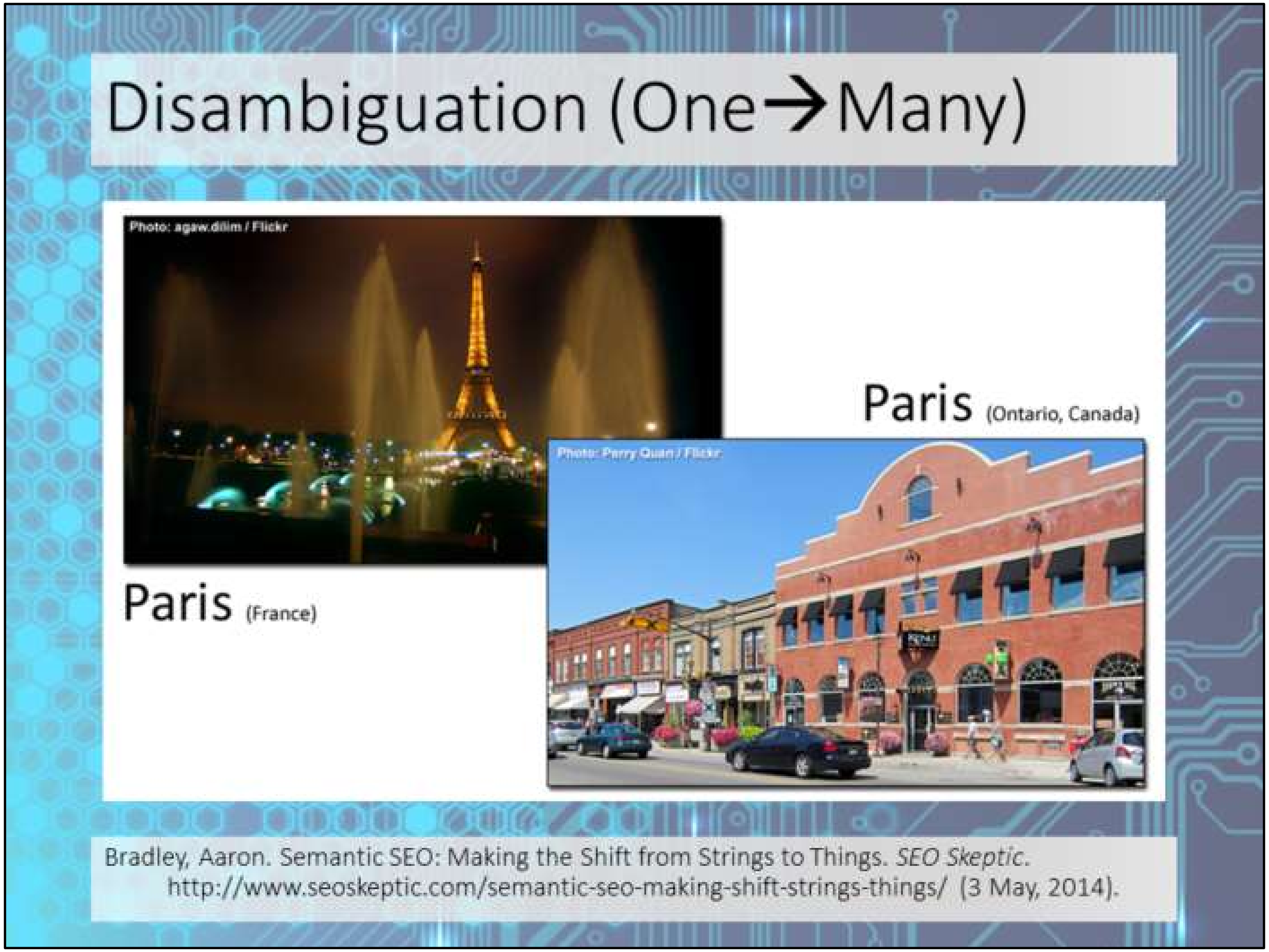
RDF relies heavily on Unique Resource Identifiers known as URIs. When you link to something, you want to know that you are linking to one specific page and that that page will be there for the forseeable future.



Disambiguation (Many→One)

Bradley, Aaron. Semantic SEO: Making the Shift from Strings to Things. *SEO Skeptic*. http://www.seoskeptic.com/semantic-seo-making-shift-strings-things/ (3 May, 2014).

This requires disambiguation.
Many names for one “Thing”



Disambiguation (One→Many)

Photo: agaw.dilim / Flickr

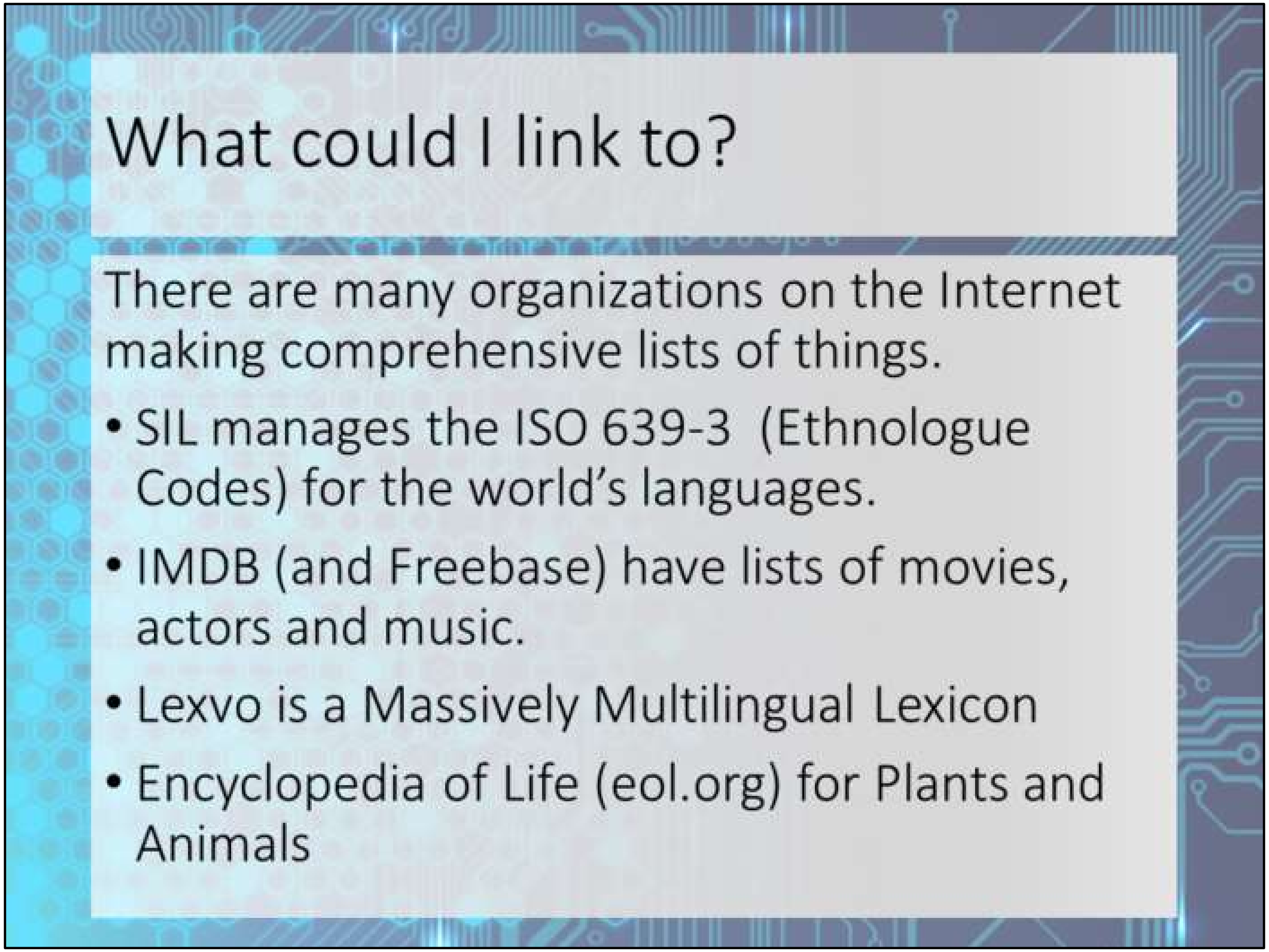
Paris (France)

Paris (Ontario, Canada)

Photo: Perry Quan / Flickr

Bradley, Aaron. Semantic SEO: Making the Shift from Strings to Things. *SEO Skeptic*. http://www.seoskeptic.com/semantic-seo-making-shift-strings-things/ (3 May, 2014).

Many things with the same name.

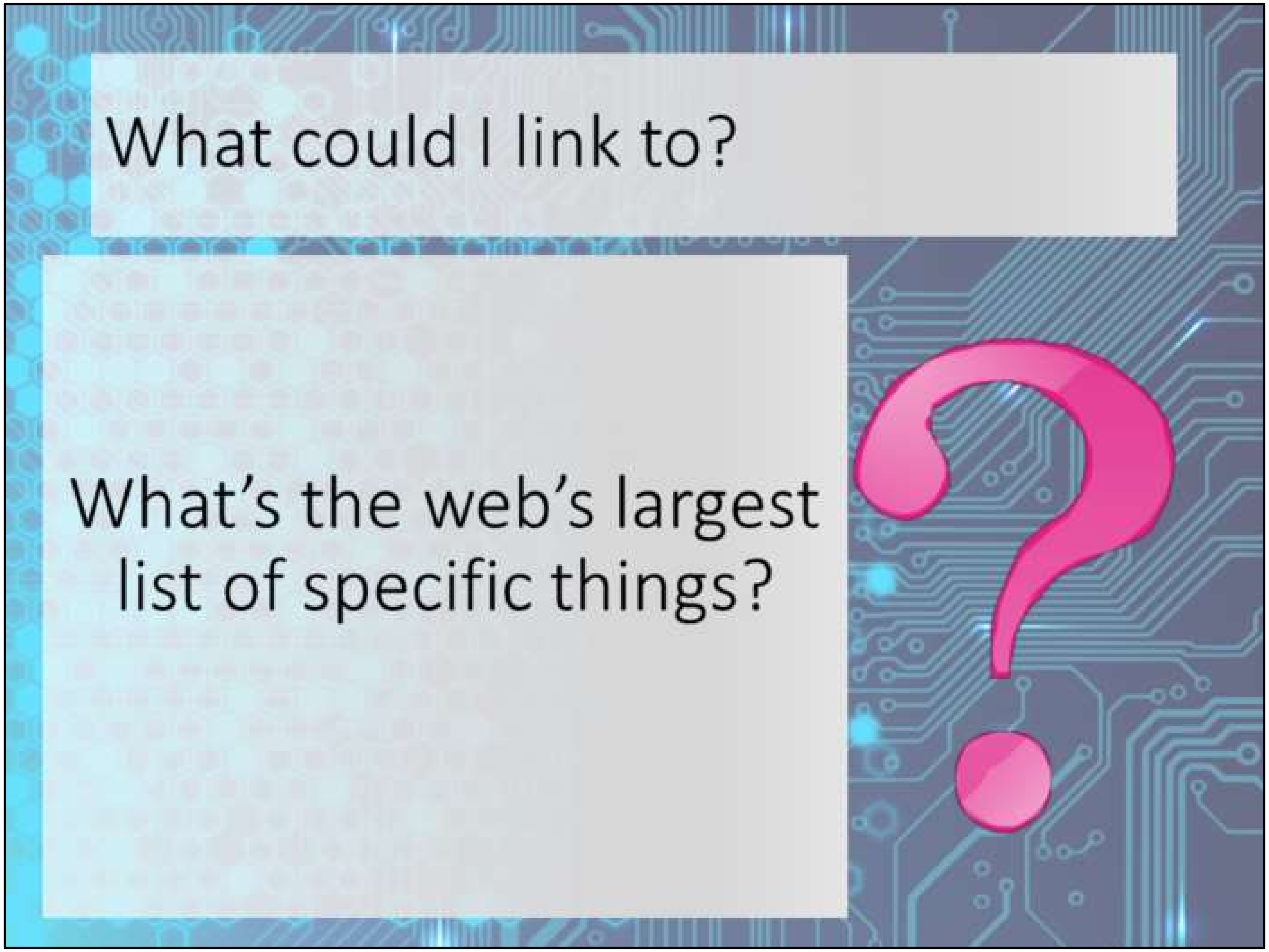


What could I link to?

There are many organizations on the Internet making comprehensive lists of things.

- SIL manages the ISO 639-3 (Ethnologue Codes) for the world's languages.
- IMDB (and Freebase) have lists of movies, actors and music.
- Lexvo is a Massively Multilingual Lexicon
- Encyclopedia of Life (eol.org) for Plants and Animals

To get around this, you can link to the specific thing on an authoritative source.

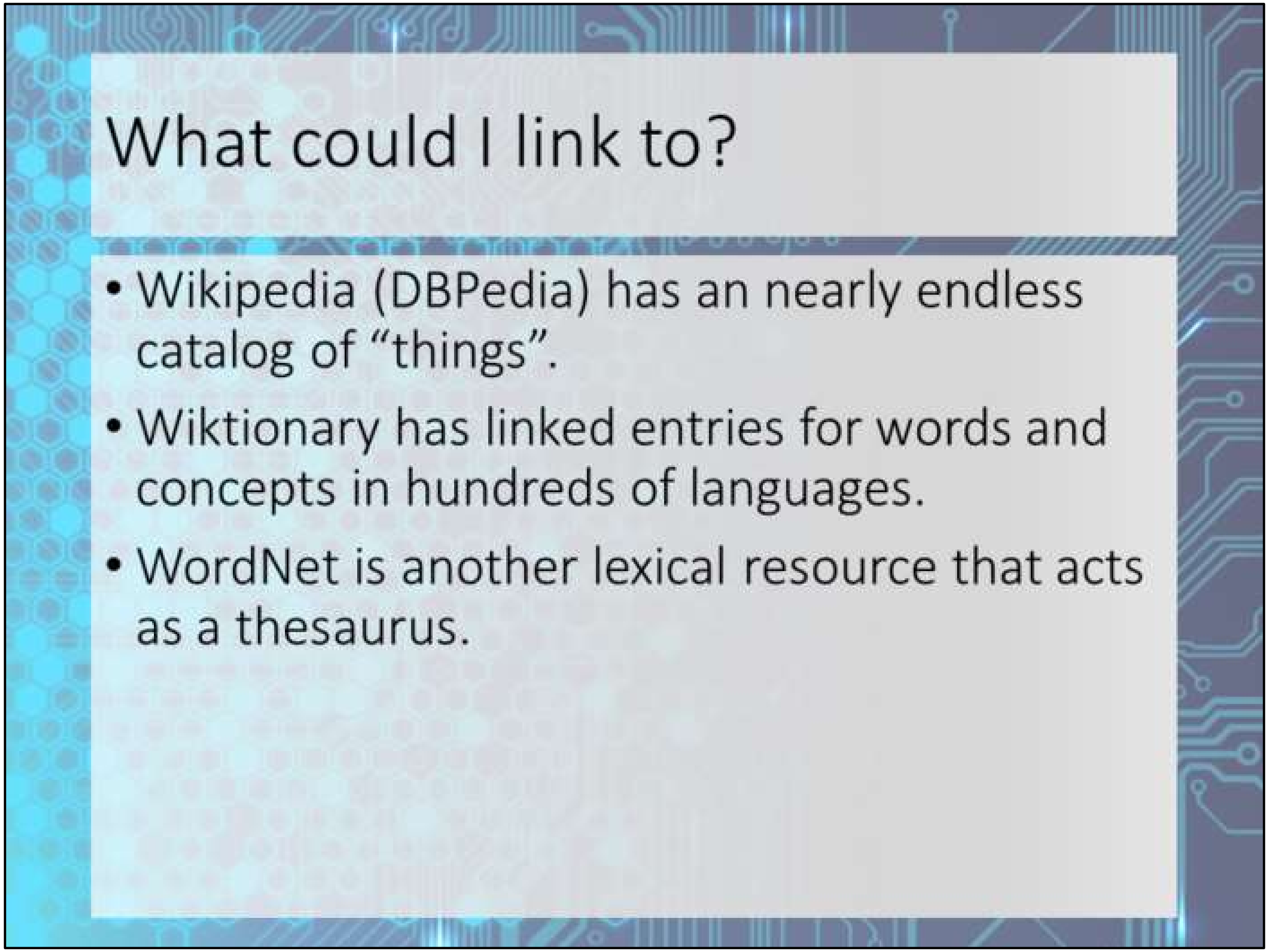


What could I link to?

What's the web's largest list of specific things?

Wikipedia and Wiktionary!

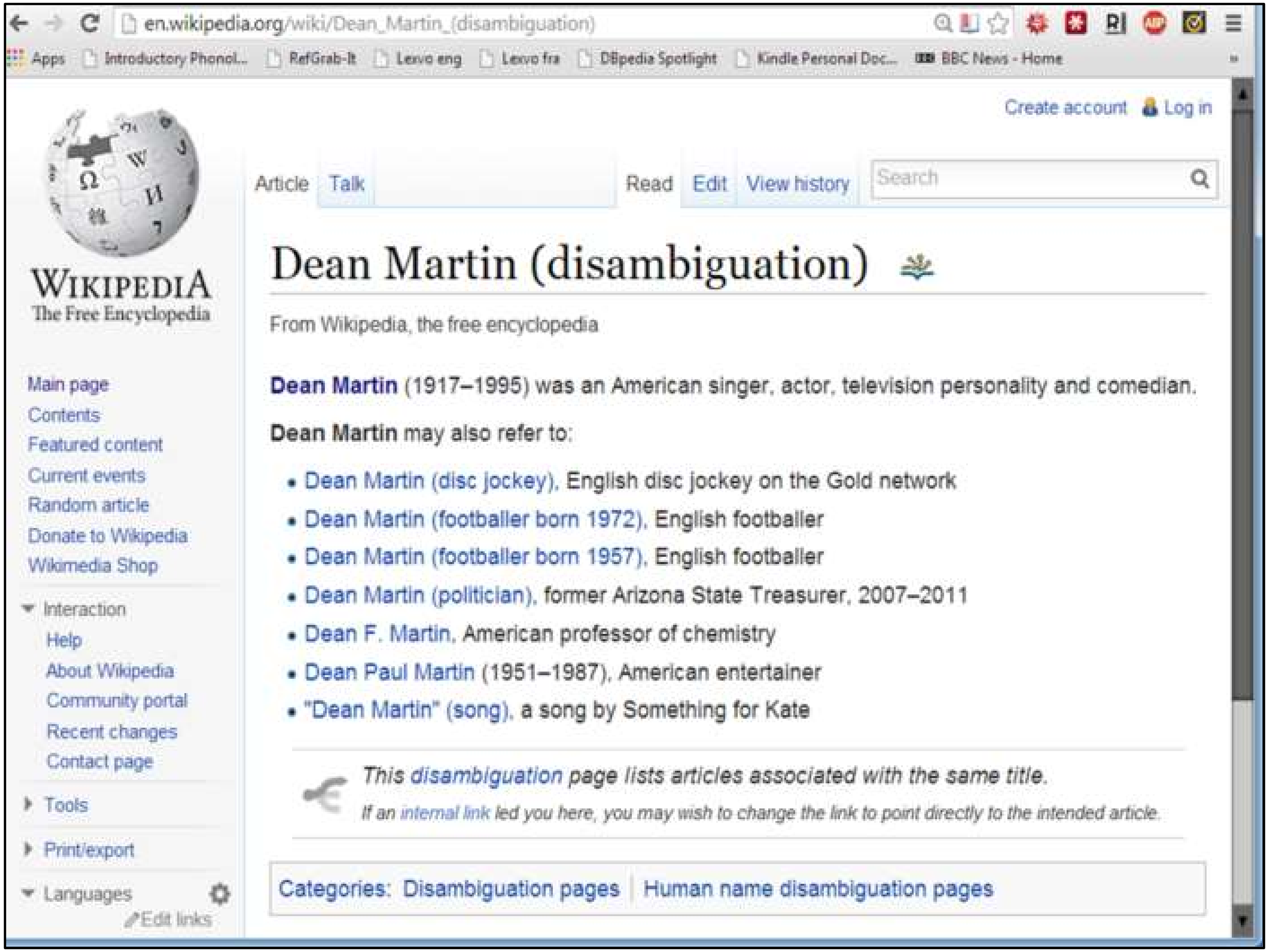
I know that Wikipedia is a **Bad Word** in the academic world. You are not saying that you agree with everything on the page, but by linking to Wikipedia you can say: I’m talking about this specific thing that exists in the real world!



What could I link to?

- Wikipedia (DBPedia) has an nearly endless catalog of “things”.
- Wiktionary has linked entries for words and concepts in hundreds of languages.
- WordNet is another lexical resource that acts as a thesaurus.

Dictionaries as Datapoints: Matthew Lee



Dean Martin (disambiguation)

From Wikipedia, the free encyclopedia

Dean Martin (1917–1995) was an American singer, actor, television personality and comedian.

Dean Martin may also refer to:

- Dean Martin (disc jockey), English disc jockey on the Gold network
- Dean Martin (footballer born 1972), English footballer
- Dean Martin (footballer born 1957), English footballer
- Dean Martin (politician), former Arizona State Treasurer, 2007–2011
- Dean F. Martin, American professor of chemistry
- Dean Paul Martin (1951–1987), American entertainer
- "Dean Martin" (song), a song by Something for Kate

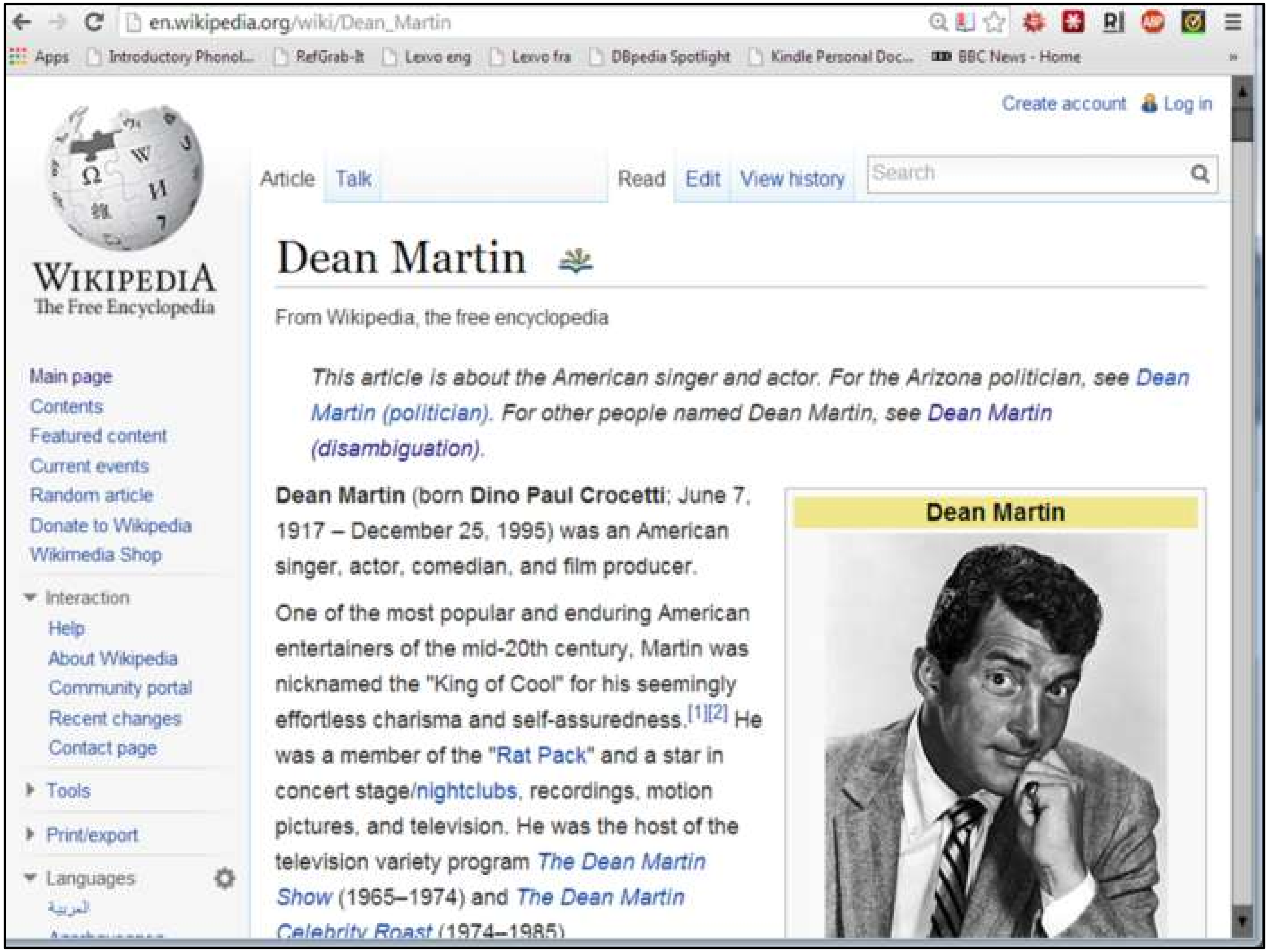
This disambiguation page lists articles associated with the same title.

If an internal link led you here, you may wish to change the link to point directly to the intended article.

Categories: Disambiguation pages | Human name disambiguation pages

Going back to our example, here we see a Wikipedia Disambiguation Page for All the Famous people named Dean Martin.

Dictionaries as Datapoints: Matthew Lee



Dean Martin

From Wikipedia, the free encyclopedia

This article is about the American singer and actor. For the Arizona politician, see Dean Martin (politician). For other people named Dean Martin, see Dean Martin (disambiguation).

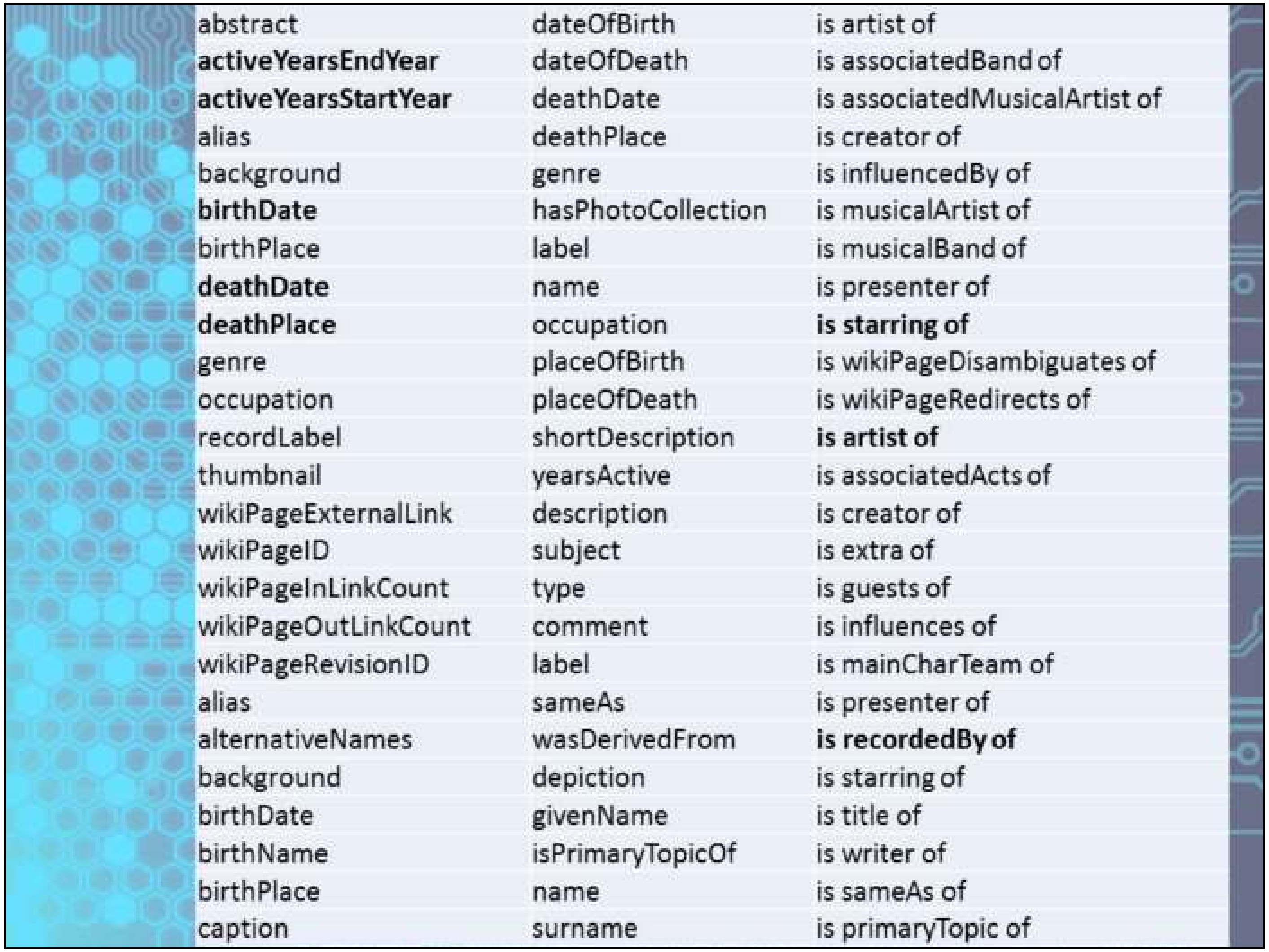
Dean Martin

Dean Martin (born **Dino Paul Crocetti**; June 7, 1917 – December 25, 1995) was an American singer, actor, comedian, and film producer.

One of the most popular and enduring American entertainers of the mid-20th century, Martin was nicknamed the "King of Cool" for his seemingly effortless charisma and self-assuredness.[1][2] He was a member of the "Rat Pack" and a star in concert stage/nightclubs, recordings, motion pictures, and television. He was the host of the television variety program *The Dean Martin Show* (1965–1974) and *The Dean Martin Celebrity Roast* (1974–1985)

And here’s the one we want.

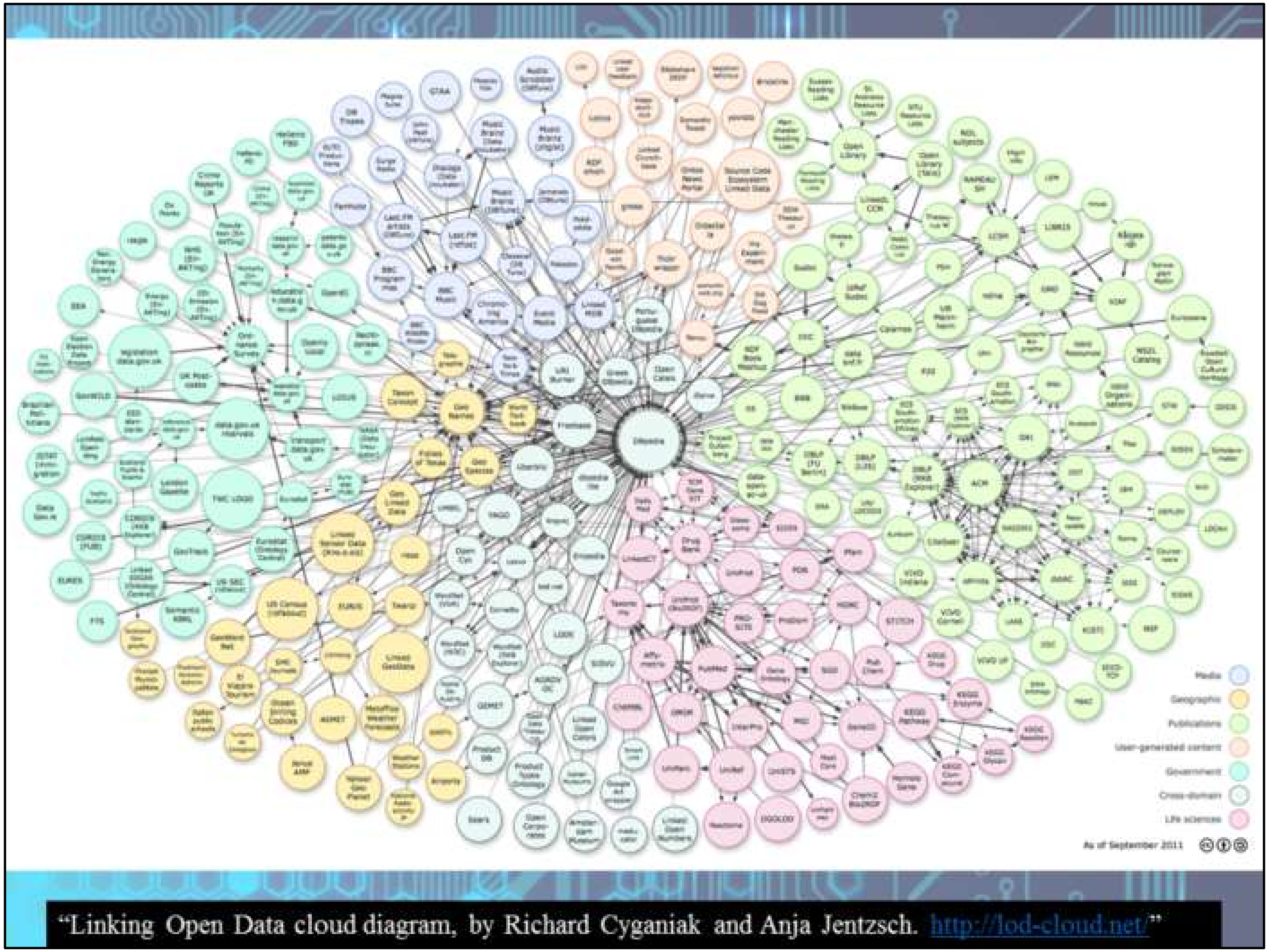
Dictionaries as Datapoints: Matthew Lee



abstract	dateOfBirth	is artist of
activeYearsEndYear	dateOfDeath	is associatedBand of
activeYearsStartYear	deathDate	is associatedMusicalArtist of
alias	deathPlace	is creator of
background	genre	is influencedBy of
birthDate	hasPhotoCollection	is musicalArtist of
birthPlace	label	is musicalBand of
deathDate	name	is presenter of
deathPlace	occupation	**is starring of**
genre	placeOfBirth	is wikiPageDisambiguates of
occupation	placeOfDeath	is wikiPageRedirects of
recordLabel	shortDescription	**is artist of**
thumbnail	yearsActive	is associatedActs of
wikiPageExternalLink	description	is creator of
wikiPageID	subject	is extra of
wikiPageInLinkCount	type	is guests of
wikiPageOutLinkCount	comment	is influences of
wikiPageRevisionID	label	is mainCharTeam of
alias	sameAs	is presenter of
alternativeNames	wasDerivedFrom	**is recordedBy of**
background	depiction	is starring of
birthDate	givenName	is title of
birthName	isPrimaryTopicOf	is writer of
birthPlace	name	is sameAs of
caption	surname	is primaryTopic of

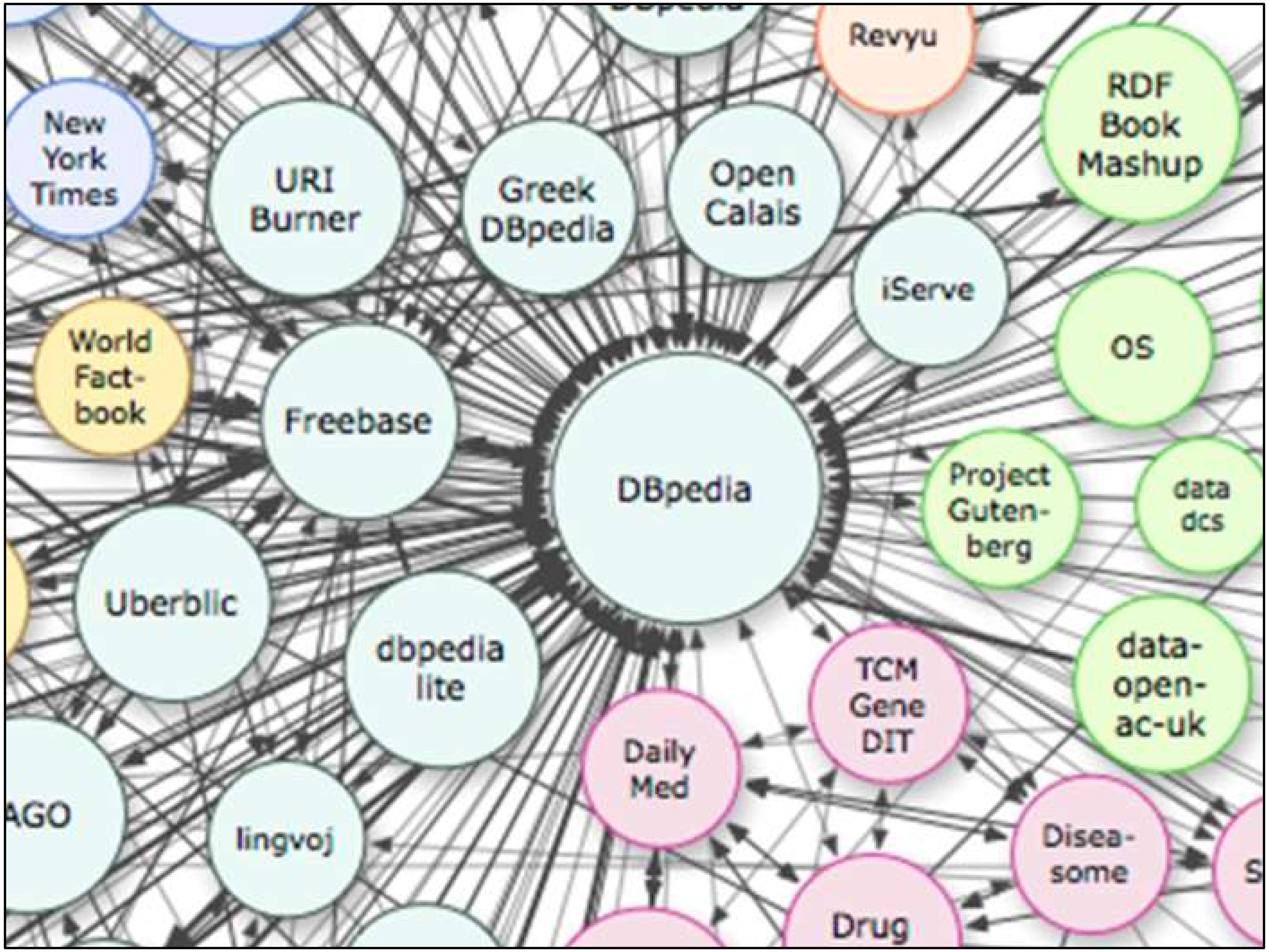
Just By linking

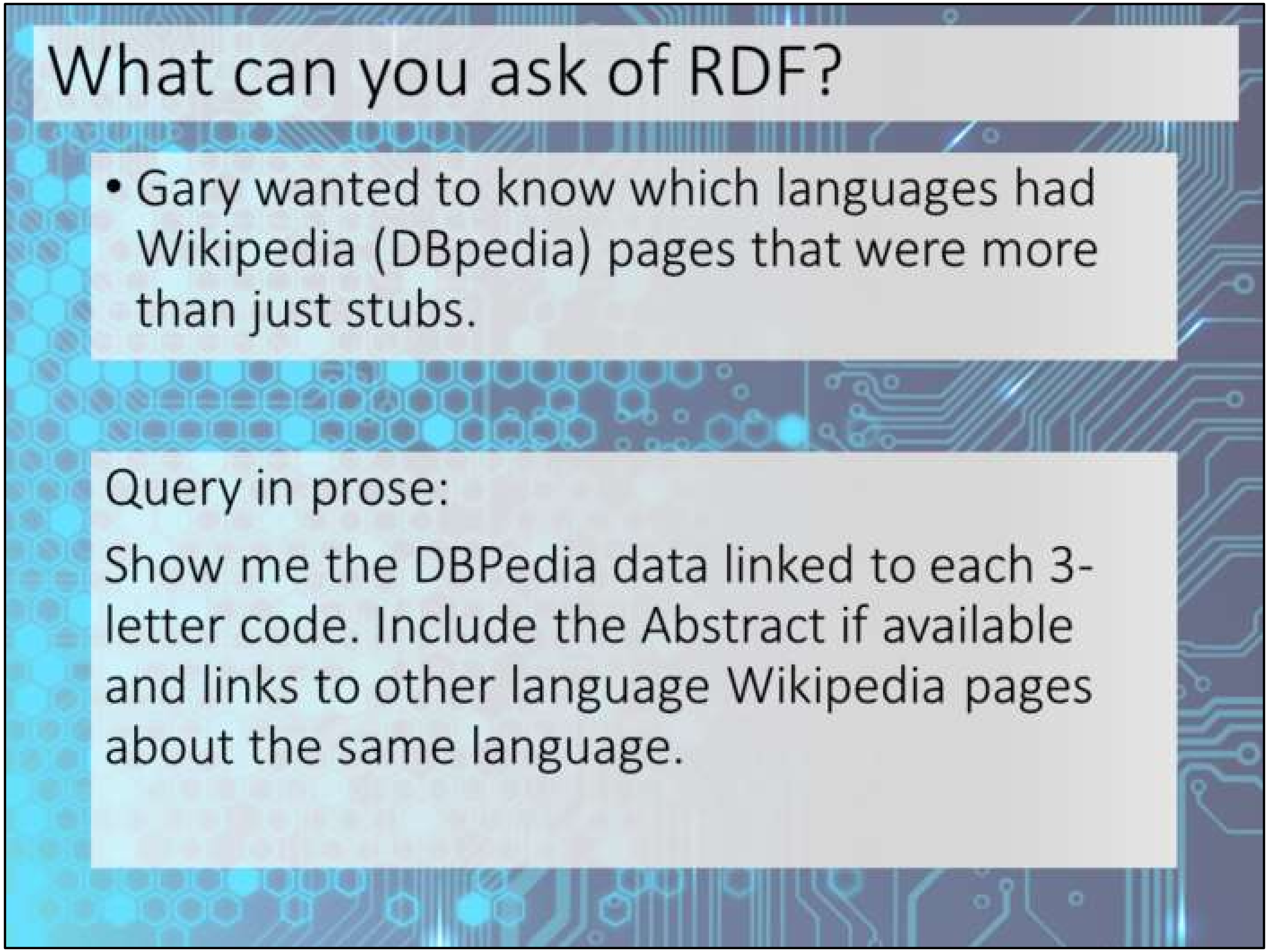
Dictionaries as Datapoints: Matthew Lee



“Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/”

Dictionaries as Datapoints: Matthew Lee





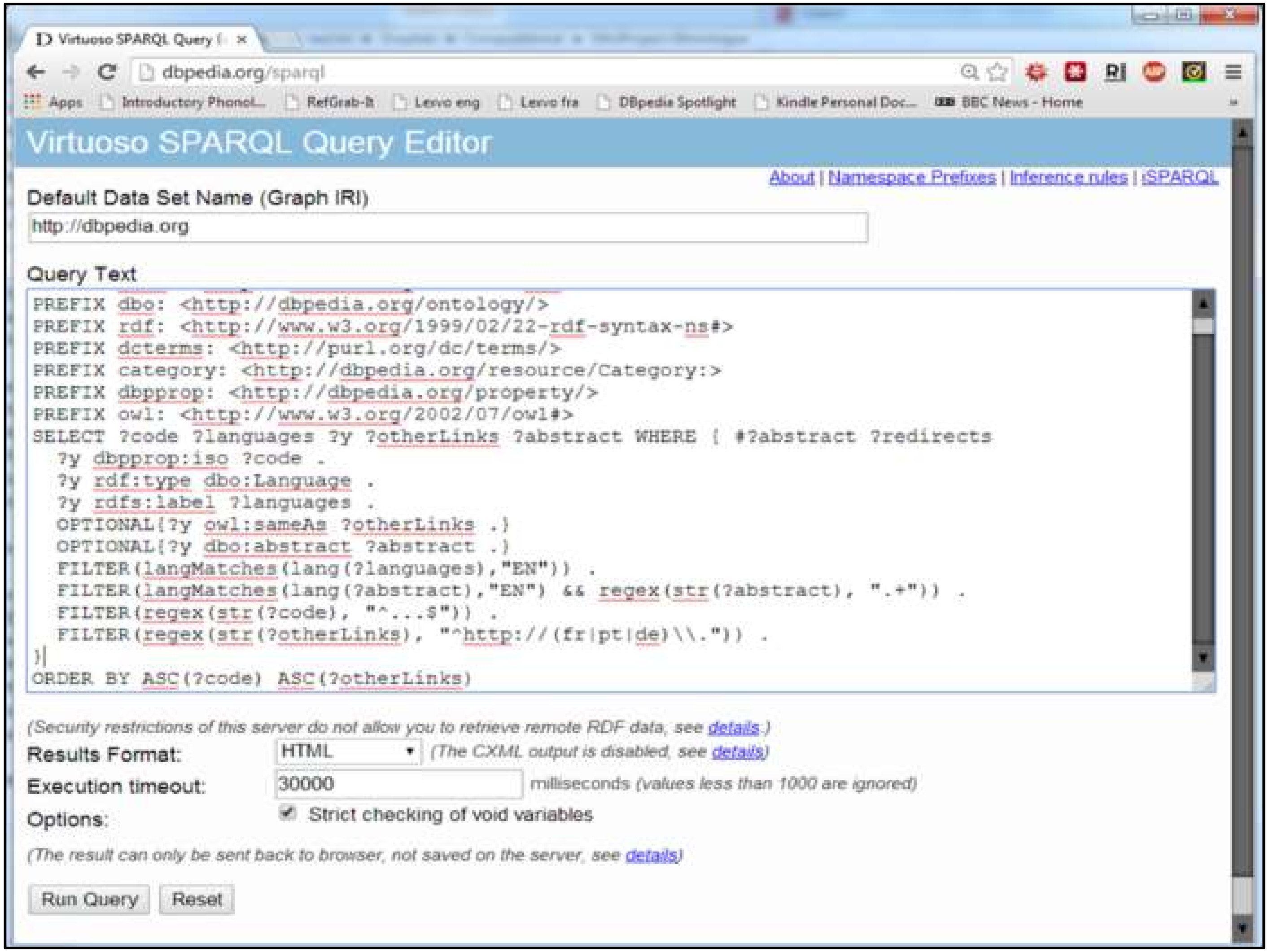
What can you ask of RDF?

- Gary wanted to know which languages had Wikipedia (DBpedia) pages that were more than just stubs.

Query in prose:

Show me the DBPedia data linked to each 3-letter code. Include the Abstract if available and links to other language Wikipedia pages about the same language.

Dictionaries as Datapoints: Matthew Lee



Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

http://dbpedia.org

Query Text

```
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX category: <http://dbpedia.org/resource/Category:>
PREFIX dbpprop: <http://dbpedia.org/property/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
SELECT ?code ?languages ?y ?otherLinks ?abstract WHERE { #?abstract ?redirects
  ?y dbpprop:iso ?code .
  ?y rdf:type dbo:Language .
  ?y rdfs:label ?languages .
  OPTIONAL{?y owl:sameAs ?otherLinks .}
  OPTIONAL{?y dbo:abstract ?abstract .}
  FILTER(langMatches(lang(?languages),"EN")) .
  FILTER(langMatches(lang(?abstract),"EN") && regex(str(?abstract), ".+")) .
  FILTER(regex(str(?code), "^...$")) .
  FILTER(regex(str(?otherLinks), "^http://(fr|pt|de)\\.")) .
}
ORDER BY ASC(?code) ASC(?otherLinks)
```

(Security restrictions of this server do not allow you to retrieve remote RDF data, see details.)

Results Format: HTML *(The CXML output is disabled, see details)*

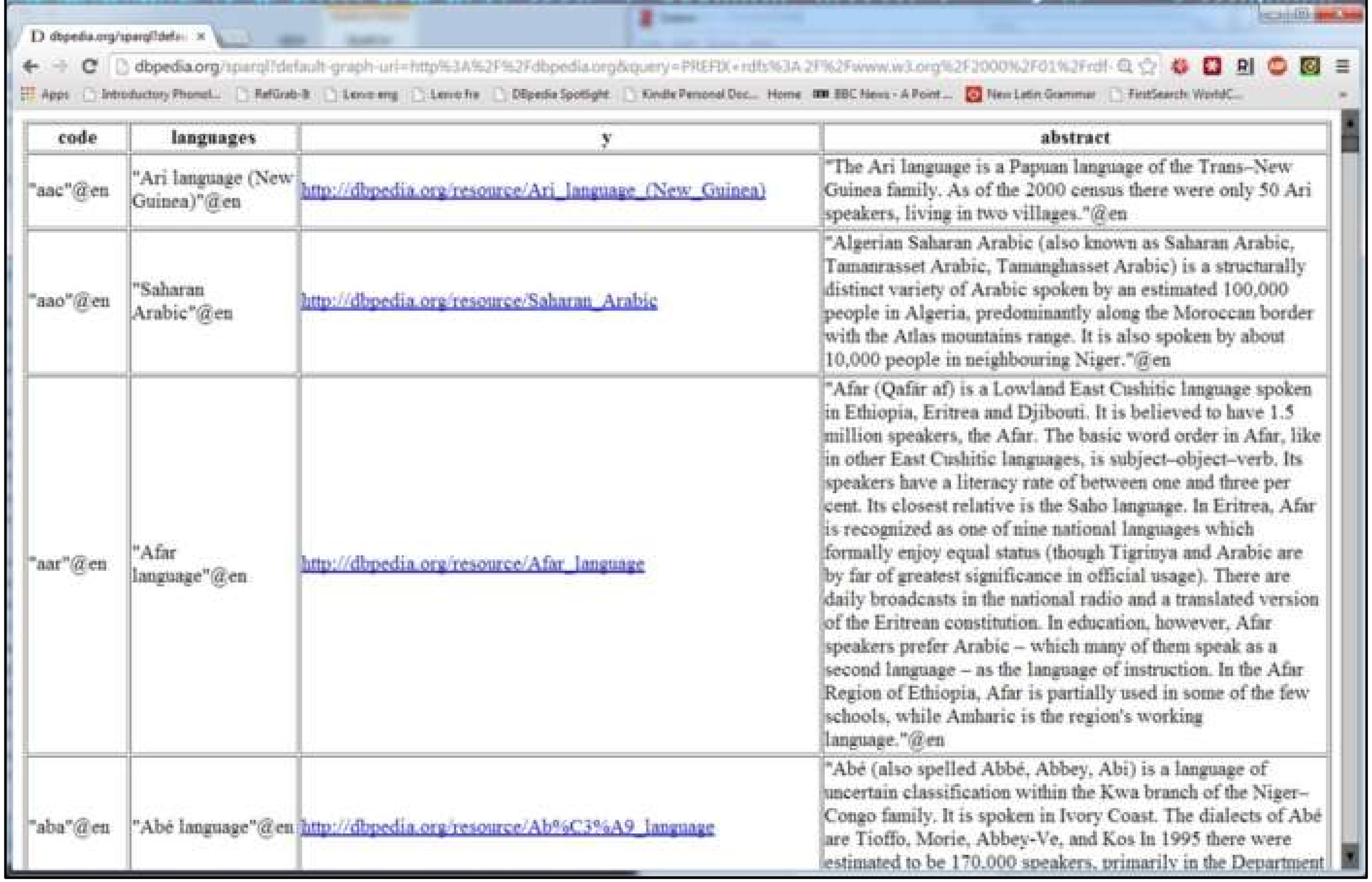
Execution timeout: 30000 milliseconds *(values less than 1000 are ignored)*

Options: ☑ Strict checking of void variables

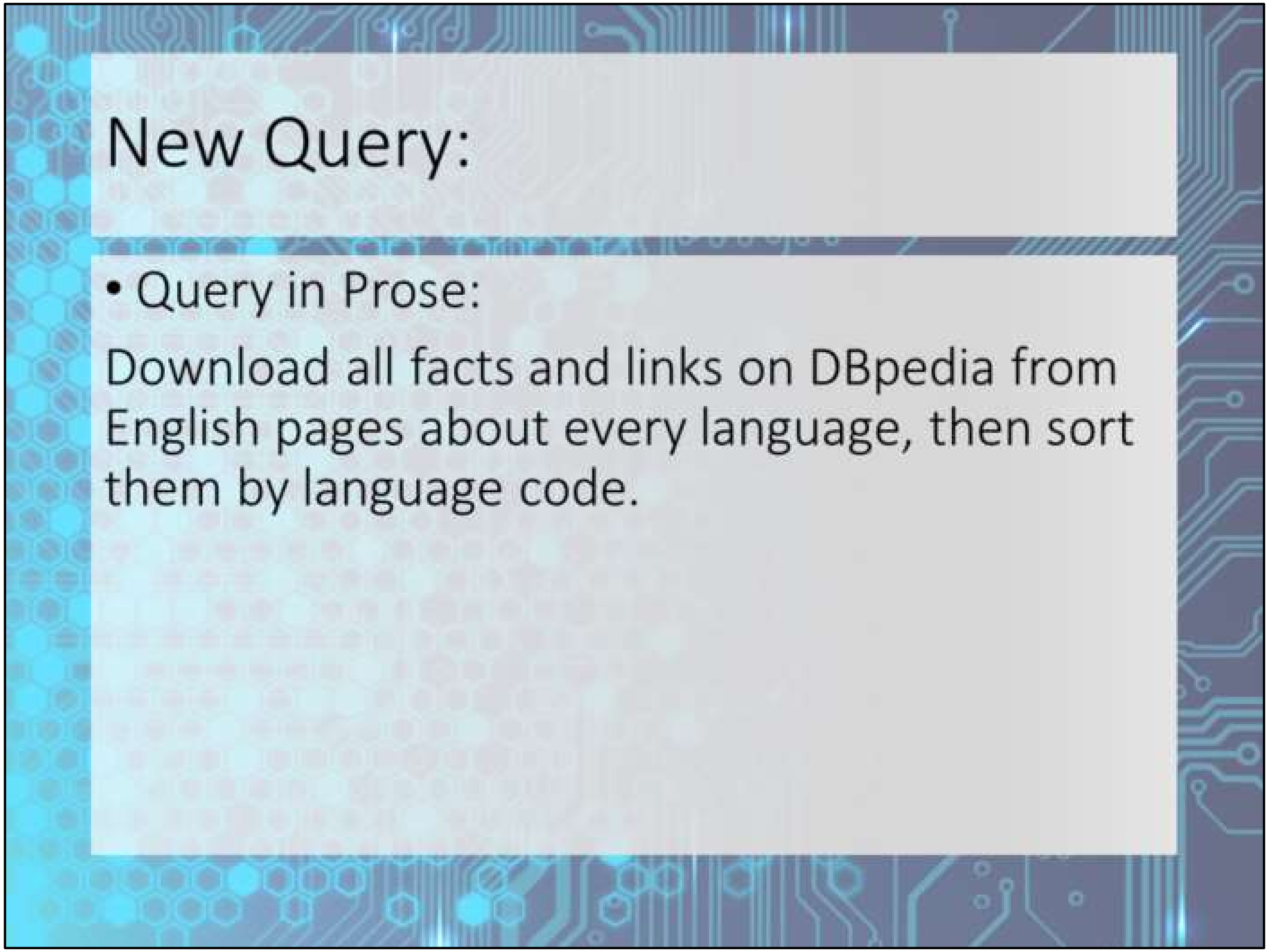
(The result can only be sent back to browser, not saved on the server, see details)

Query in SPARQL

Results:



code	languages	y	abstract
"aac"@en	"Ari language (New Guinea)"@en	http://dbpedia.org/resource/Ari_language_(New_Guinea)	"The Ari language is a Papuan language of the Trans–New Guinea family. As of the 2000 census there were only 50 Ari speakers, living in two villages."@en
"aao"@en	"Saharan Arabic"@en	http://dbpedia.org/resource/Saharan_Arabic	"Algerian Saharan Arabic (also known as Saharan Arabic, Tamanrasset Arabic, Tamanghasset Arabic) is a structurally distinct variety of Arabic spoken by an estimated 100,000 people in Algeria, predominantly along the Moroccan border with the Atlas mountains range. It is also spoken by about 10,000 people in neighbouring Niger."@en
"aar"@en	"Afar language"@en	http://dbpedia.org/resource/Afar_language	"Afar (Qafár af) is a Lowland East Cushitic language spoken in Ethiopia, Eritrea and Djibouti. It is believed to have 1.5 million speakers, the Afar. The basic word order in Afar, like in other East Cushitic languages, is subject–object–verb. Its speakers have a literacy rate of between one and three per cent. Its closest relative is the Saho language. In Eritrea, Afar is recognized as one of nine national languages which formally enjoy equal status (though Tigrinya and Arabic are by far of greatest significance in official usage). There are daily broadcasts in the national radio and a translated version of the Eritrean constitution. In education, however, Afar speakers prefer Arabic – which many of them speak as a second language – as the language of instruction. In the Afar Region of Ethiopia, Afar is partially used in some of the few schools, while Amharic is the region's working language."@en
"aba"@en	"Abé language"@en	http://dbpedia.org/resource/Ab%C3%A9_language	"Abé (also spelled Abbé, Abbey, Abi) is a language of uncertain classification within the Kwa branch of the Niger–Congo family. It is spoken in Ivory Coast. The dialects of Abé are Tioffo, Morie, Abbey-Ve, and Kos In 1995 there were estimated to be 170,000 speakers, primarily in the Department

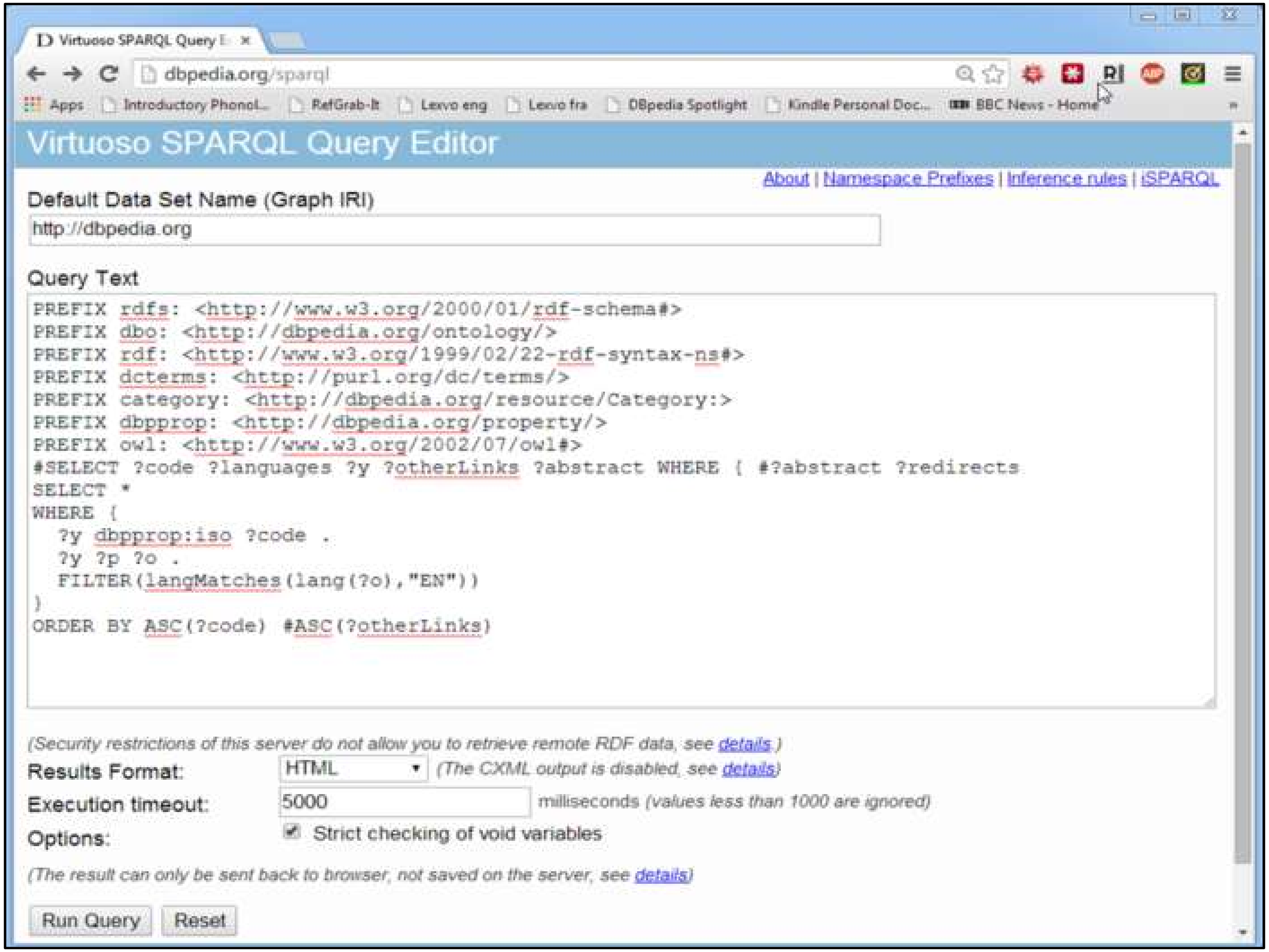


New Query:

- Query in Prose:

Download all facts and links on DBpedia from English pages about every language, then sort them by language code.

Dictionaries as Datapoints: Matthew Lee



Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

http://dbpedia.org

Query Text

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dbo: <http://dbpedia.org/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX category: <http://dbpedia.org/resource/Category:>
PREFIX dbpprop: <http://dbpedia.org/property/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
#SELECT ?code ?languages ?y ?otherLinks ?abstract WHERE { #?abstract ?redirects
SELECT *
WHERE {
  ?y dbpprop:iso ?code .
  ?y ?p ?o .
  FILTER(langMatches(lang(?o),"EN"))
}
ORDER BY ASC(?code) #ASC(?otherLinks)
```

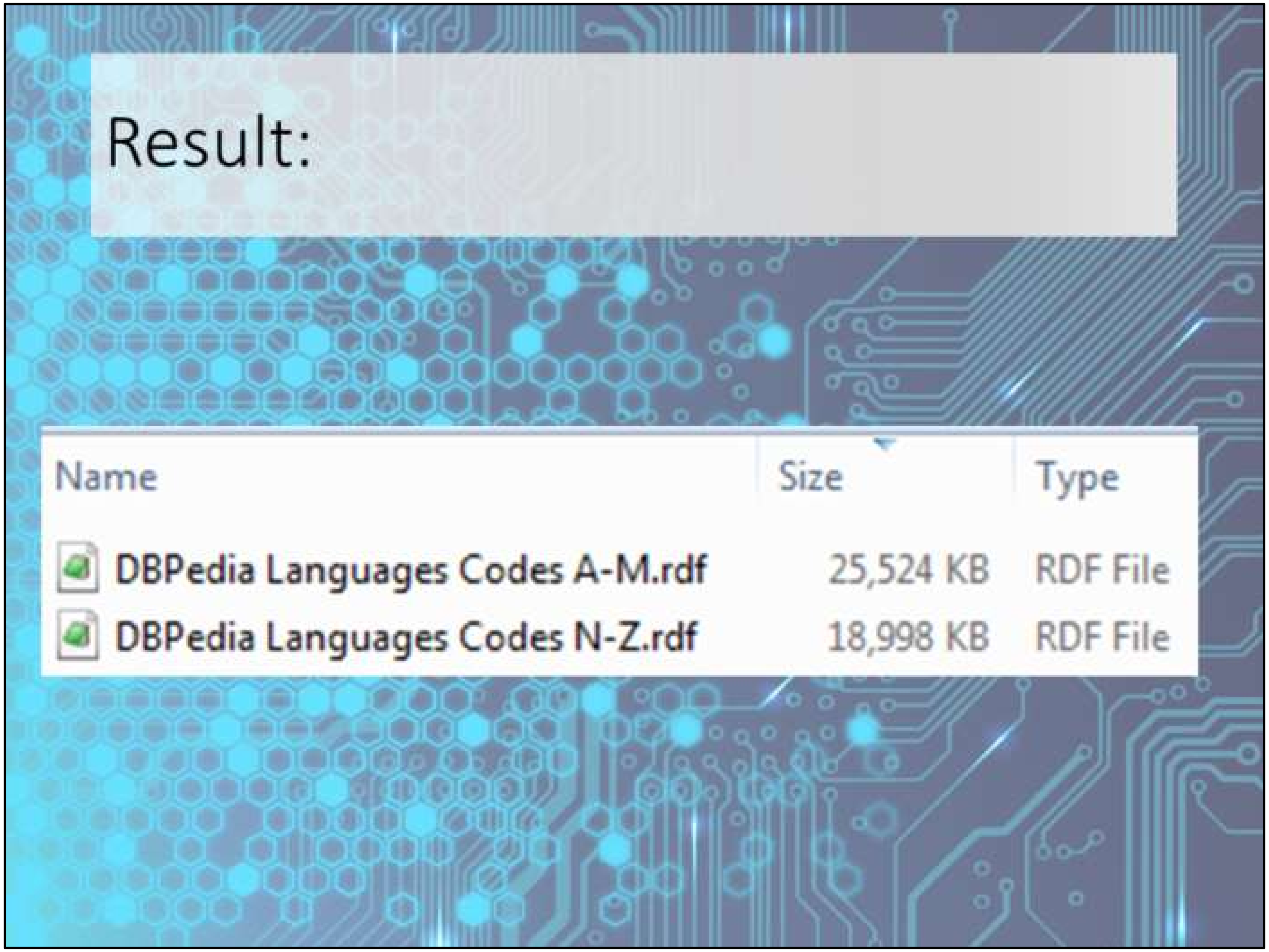
(Security restrictions of this server do not allow you to retrieve remote RDF data, see details.)

Results Format: HTML ▾ *(The CXML output is disabled, see details)*

Execution timeout: 5000 milliseconds *(values less than 1000 are ignored)*

Options: ☑ Strict checking of void variables

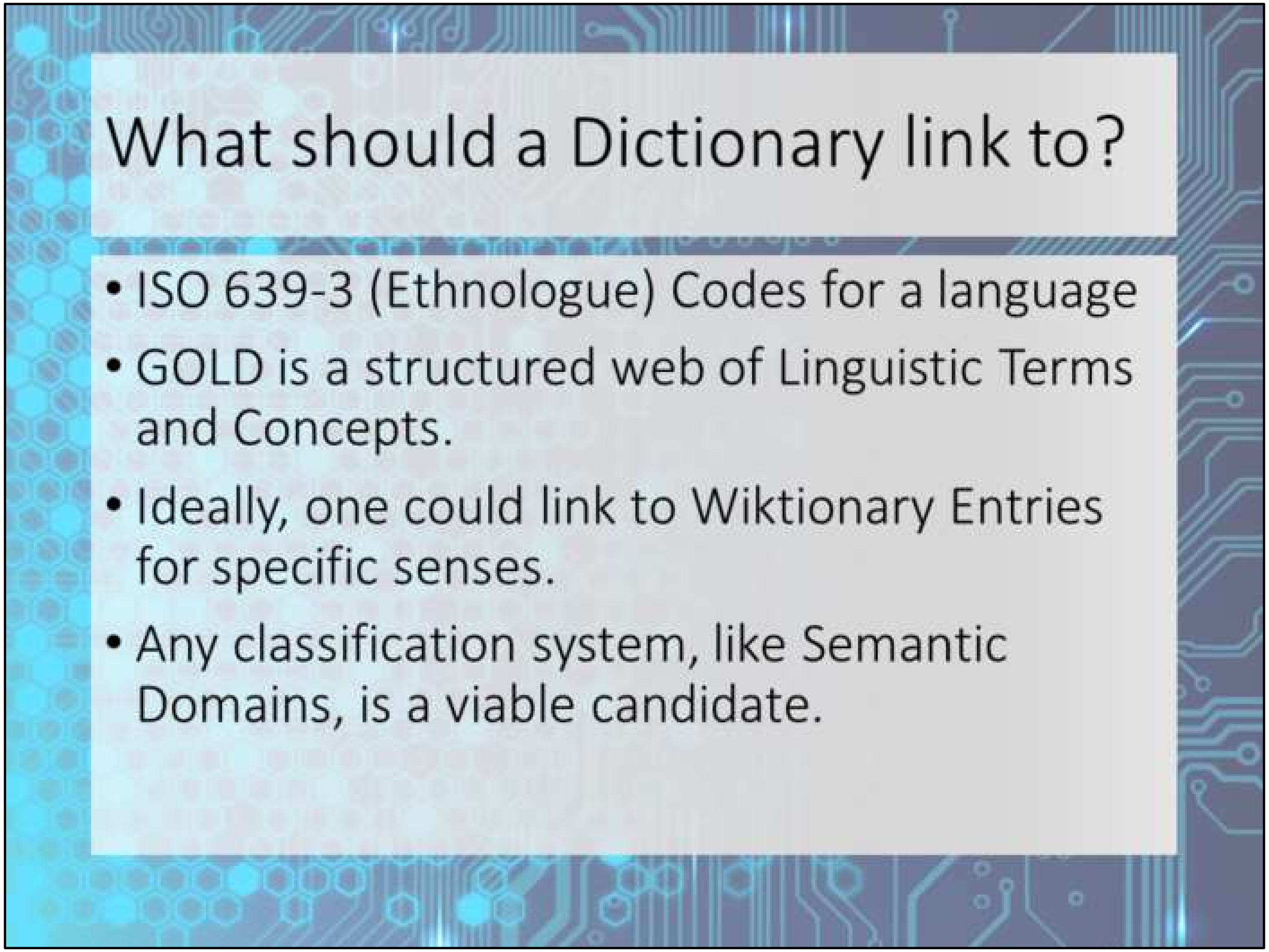
(The result can only be sent back to browser, not saved on the server, see details)



Result:

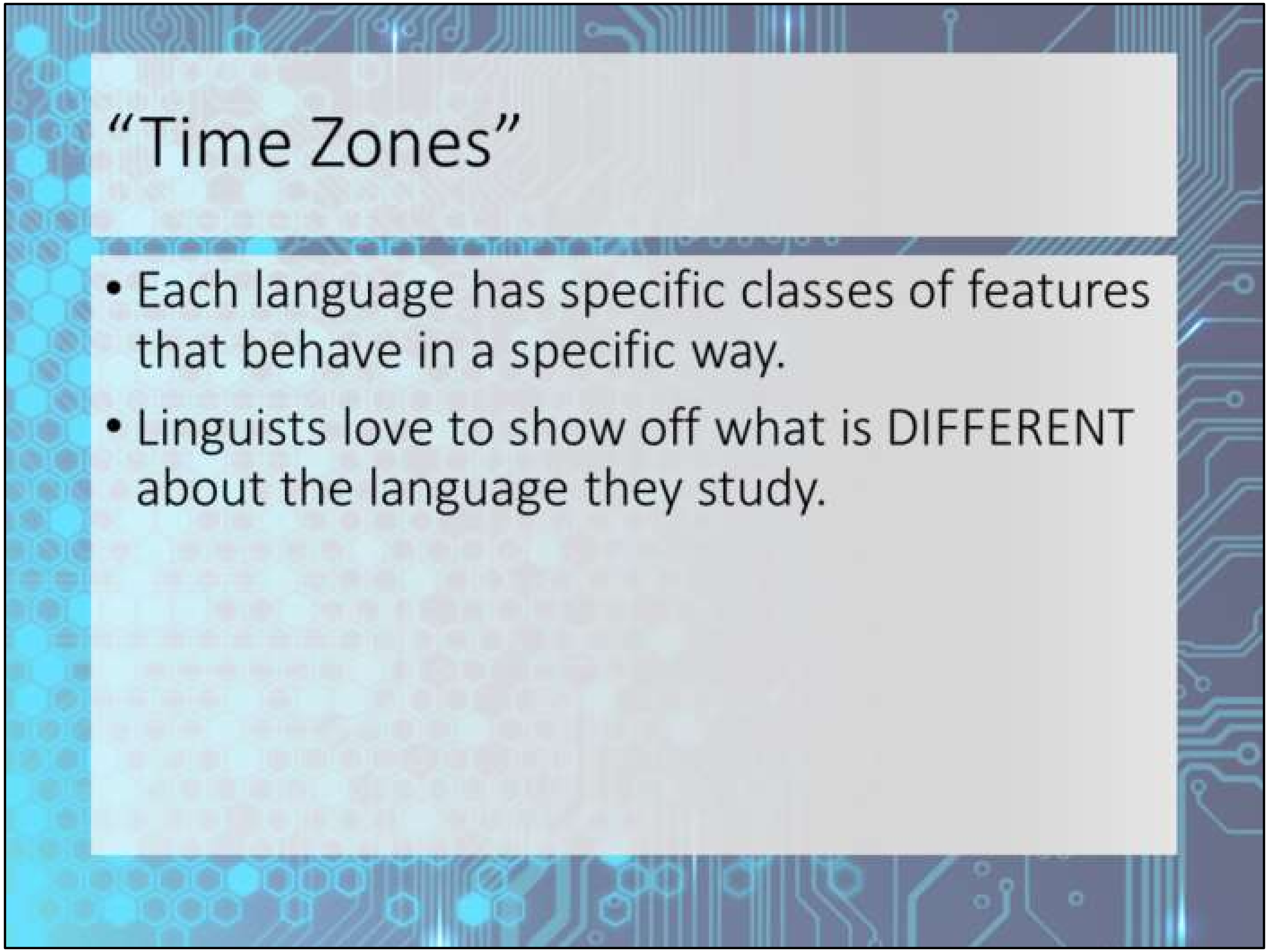
Name	Size	Type
DBPedia Languages Codes A-M.rdf	25,524 KB	RDF File
DBPedia Languages Codes N-Z.rdf	18,998 KB	RDF File

Links for Lexicons



What should a Dictionary link to?

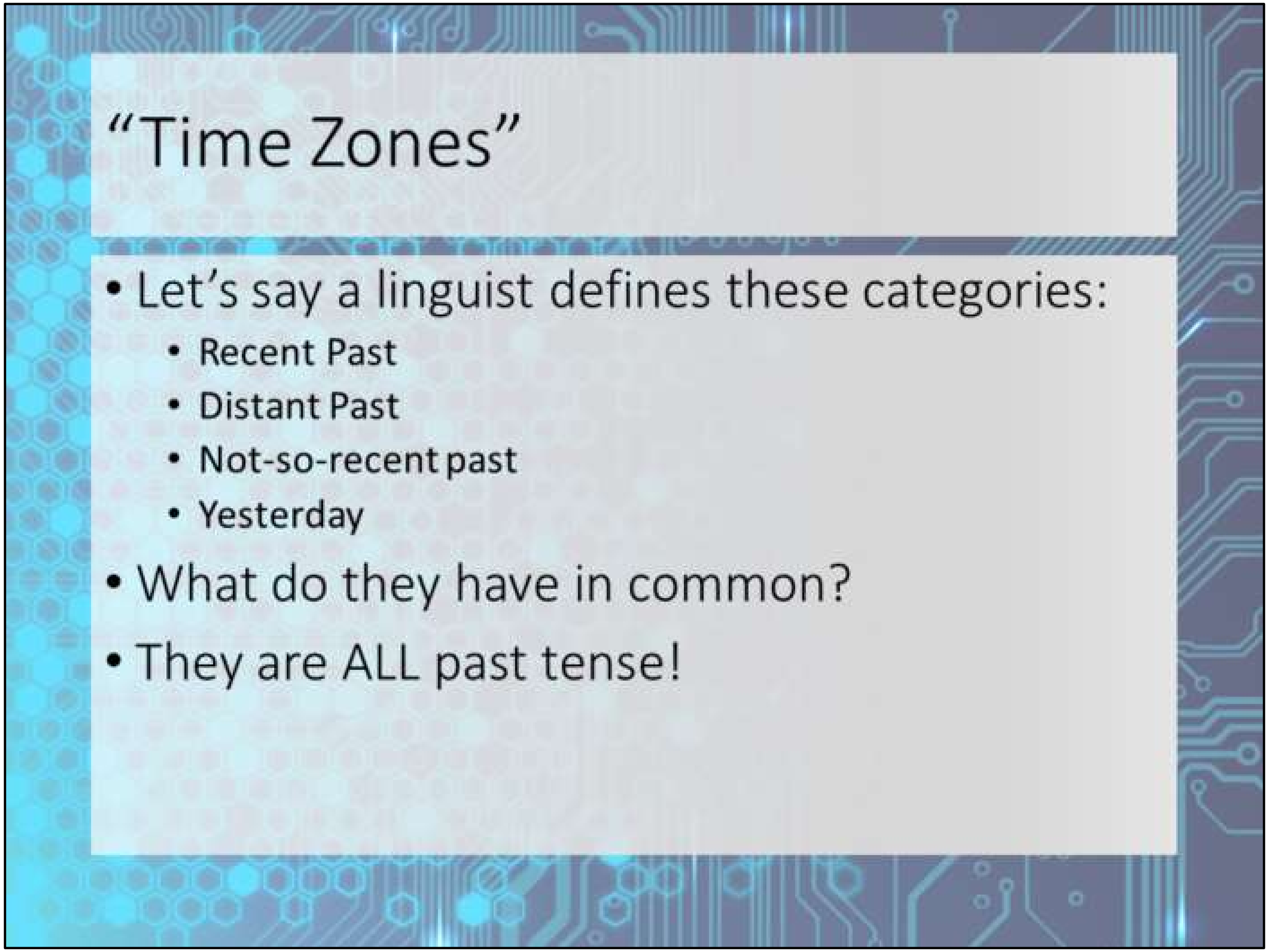
- ISO 639-3 (Ethnologue) Codes for a language
- GOLD is a structured web of Linguistic Terms and Concepts.
- Ideally, one could link to Wiktionary Entries for specific senses.
- Any classification system, like Semantic Domains, is a viable candidate.



“Time Zones”

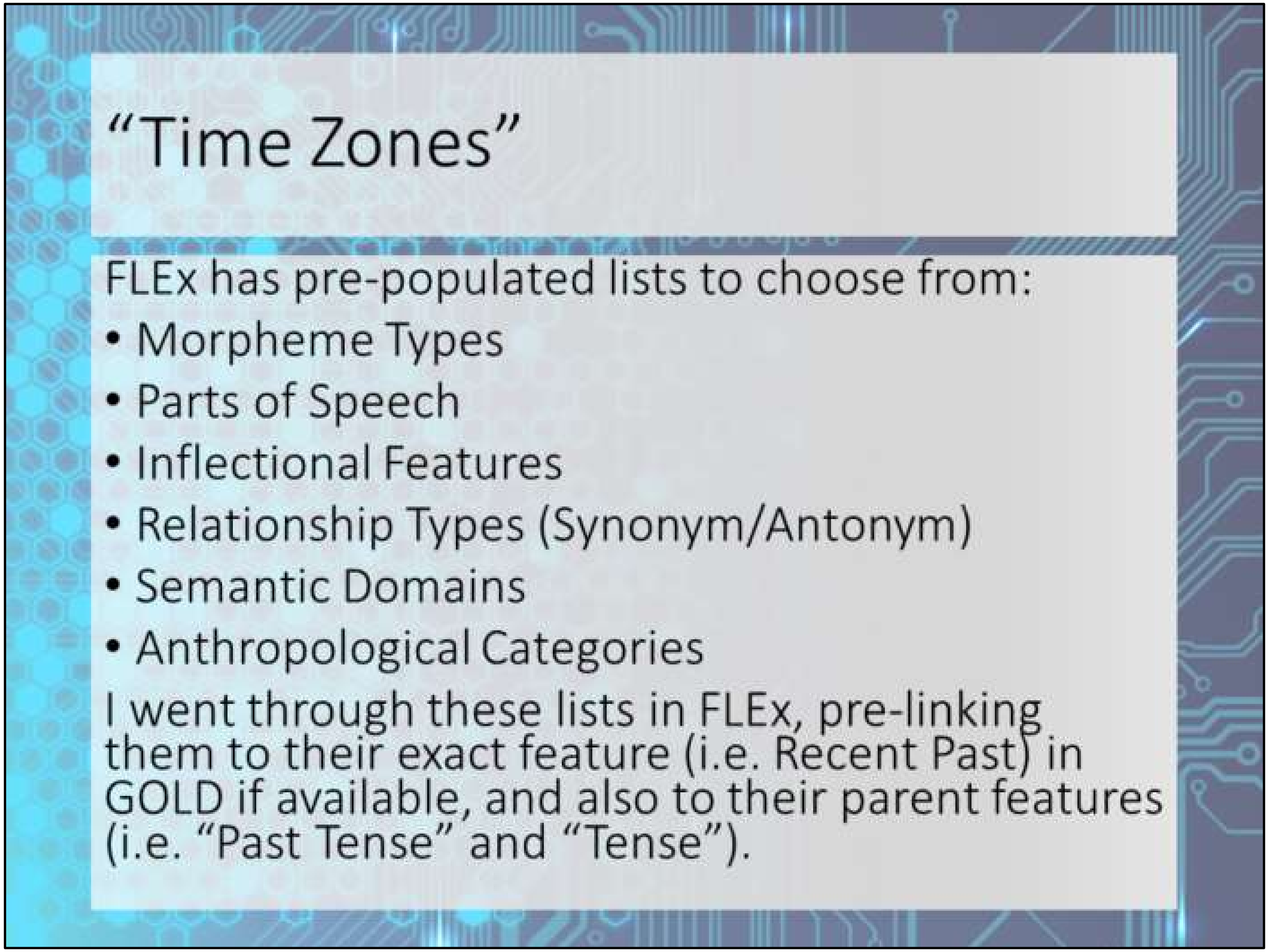
- Each language has specific classes of features that behave in a specific way.
- Linguists love to show off what is DIFFERENT about the language they study.

…so that they can become rich and famous “in the linguistic world”.



“Time Zones”

- Let’s say a linguist defines these categories:
  - Recent Past
  - Distant Past
  - Not-so-recent past
  - Yesterday
- What do they have in common?
- They are ALL past tense!

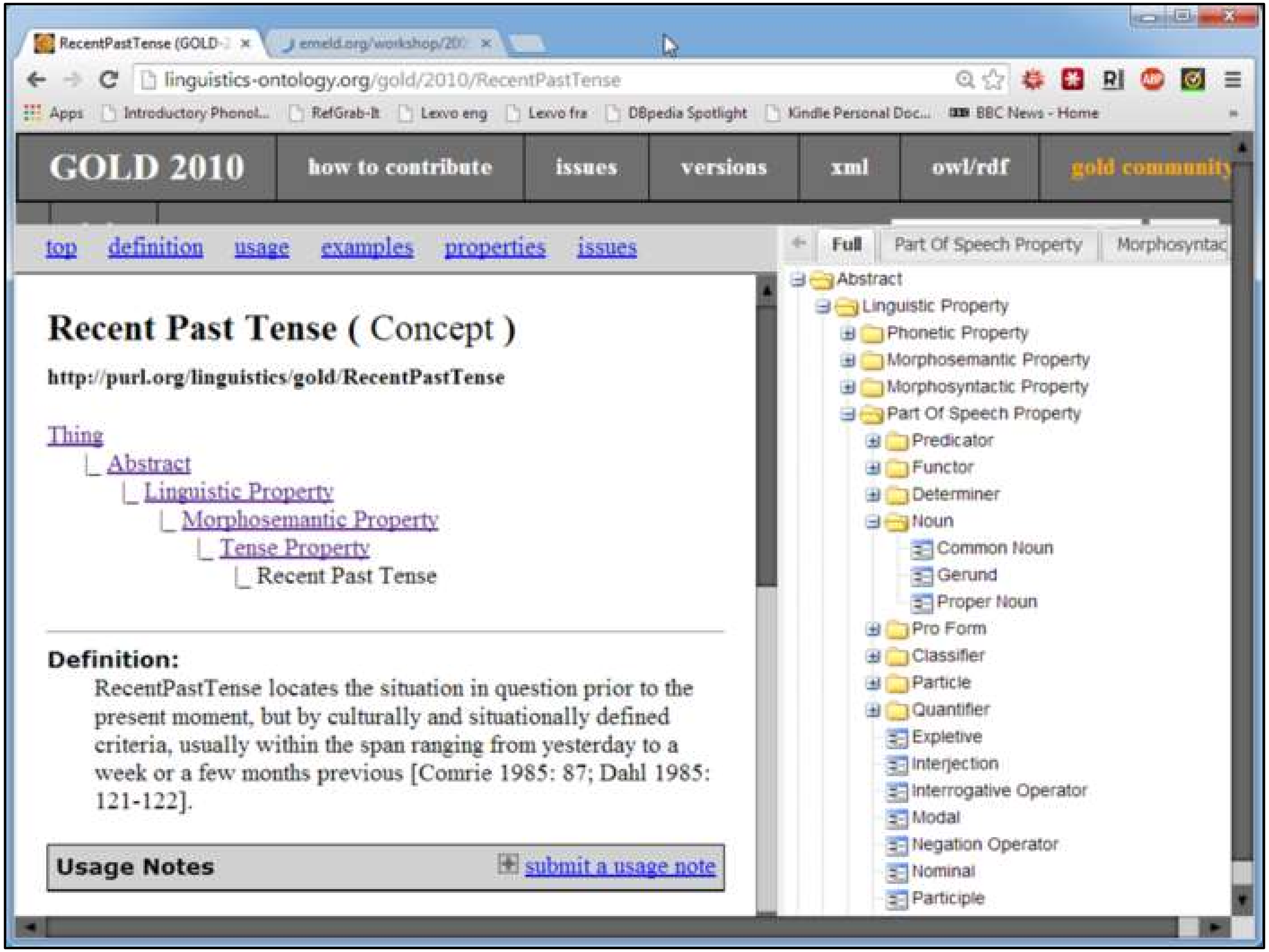


“Time Zones”

FLEx has pre-populated lists to choose from:

- Morpheme Types
- Parts of Speech
- Inflectional Features
- Relationship Types (Synonym/Antonym)
- Semantic Domains
- Anthropological Categories

I went through these lists in FLEx, pre-linking them to their exact feature (i.e. Recent Past) in GOLD if available, and also to their parent features (i.e. “Past Tense” and “Tense”).



Recent Past Tense (Concept)

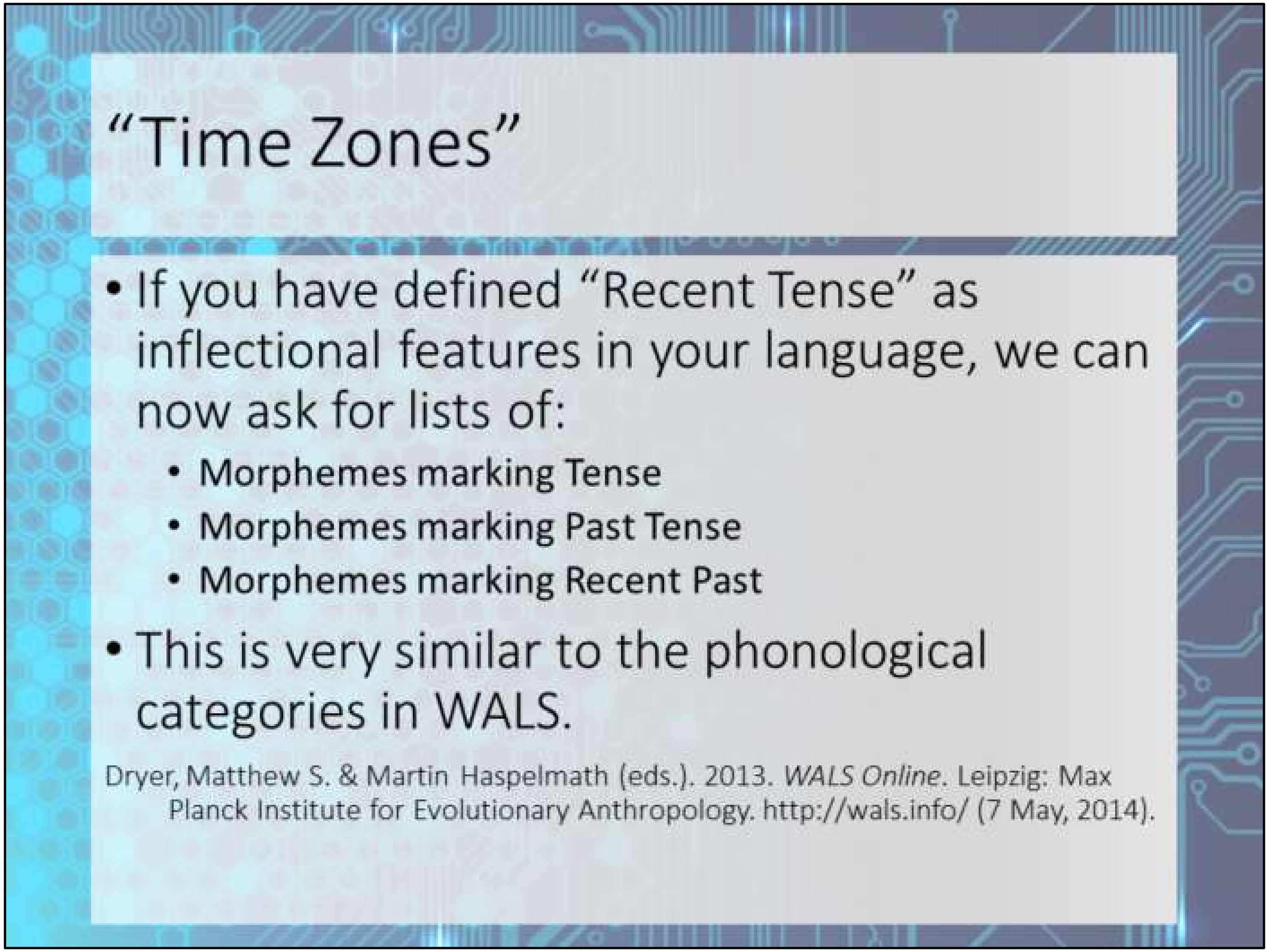
http://purl.org/linguistics/gold/RecentPastTense

Thing
  |_ Abstract
    |_ Linguistic Property
      |_ Morphosemantic Property
        |_ Tense Property
          |_ Recent Past Tense

Definition:

RecentPastTense locates the situation in question prior to the present moment, but by culturally and situationally defined criteria, usually within the span ranging from yesterday to a week or a few months previous [Comrie 1985: 87; Dahl 1985: 121-122].

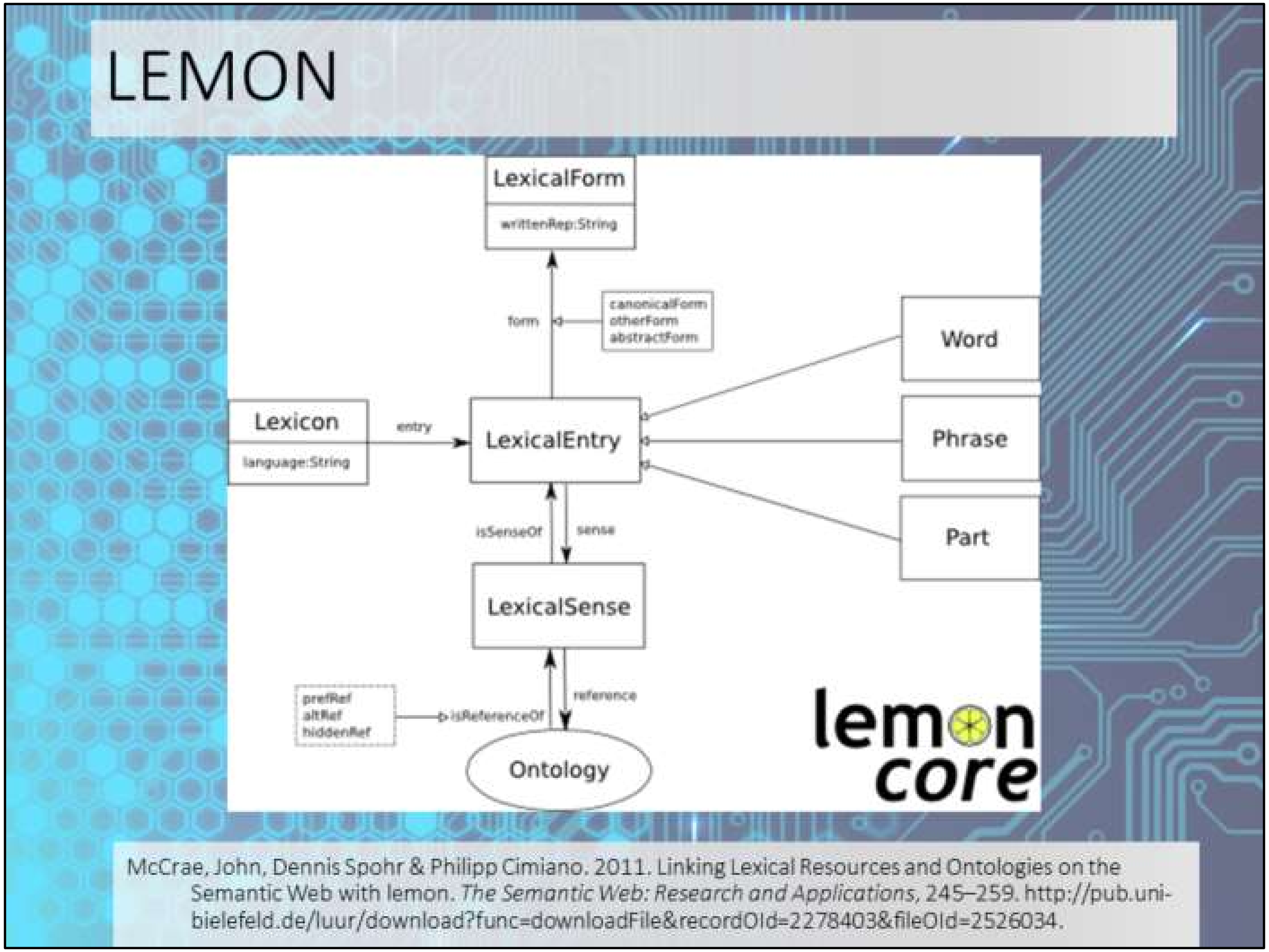
Usage Notes submit a usage note



"Time Zones"

- If you have defined "Recent Tense" as inflectional features in your language, we can now ask for lists of:
  - Morphemes marking Tense
  - Morphemes marking Past Tense
  - Morphemes marking Recent Past
- This is very similar to the phonological categories in WALS.

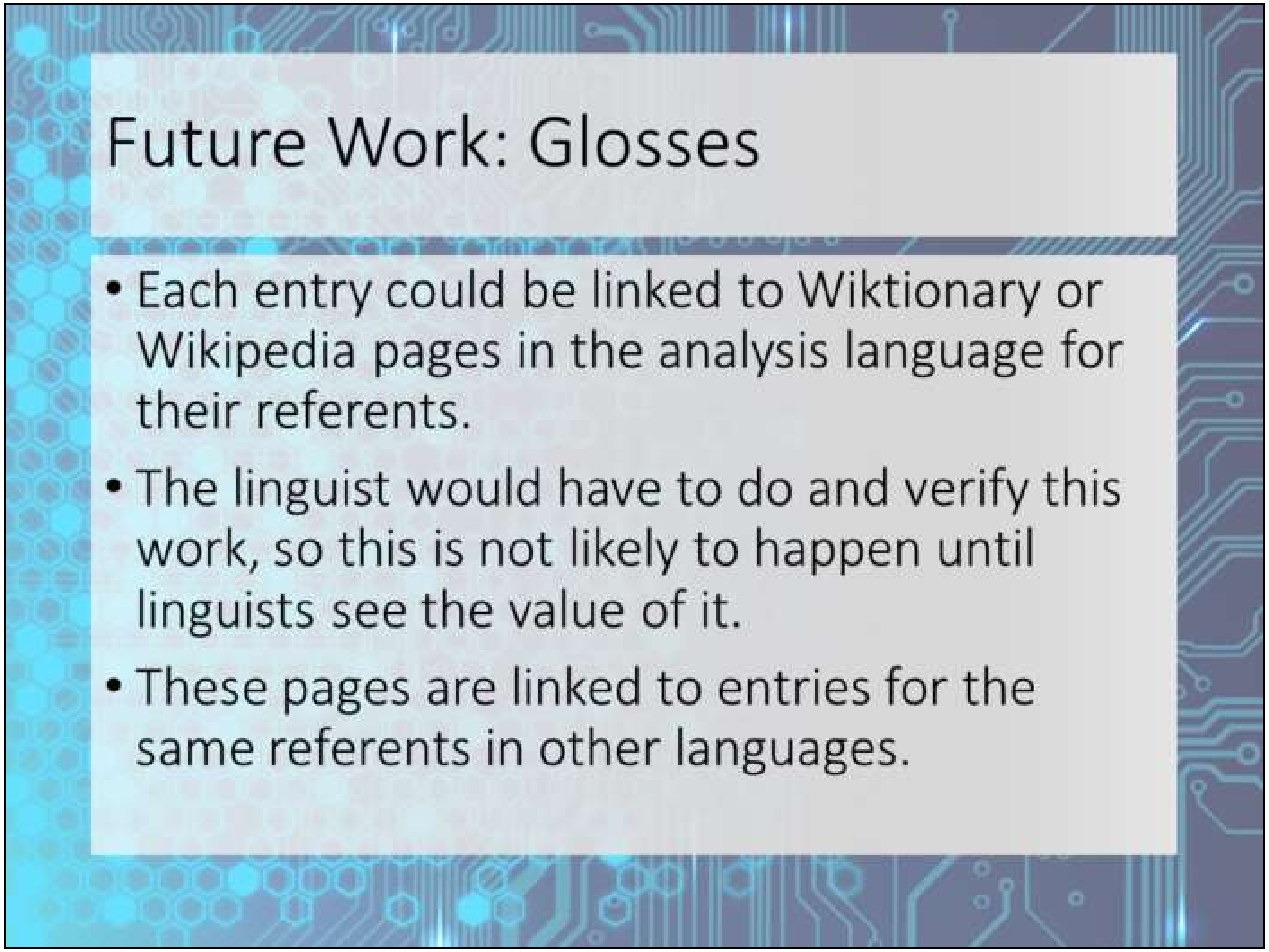
Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://wals.info/ (7 May, 2014).



LEMON

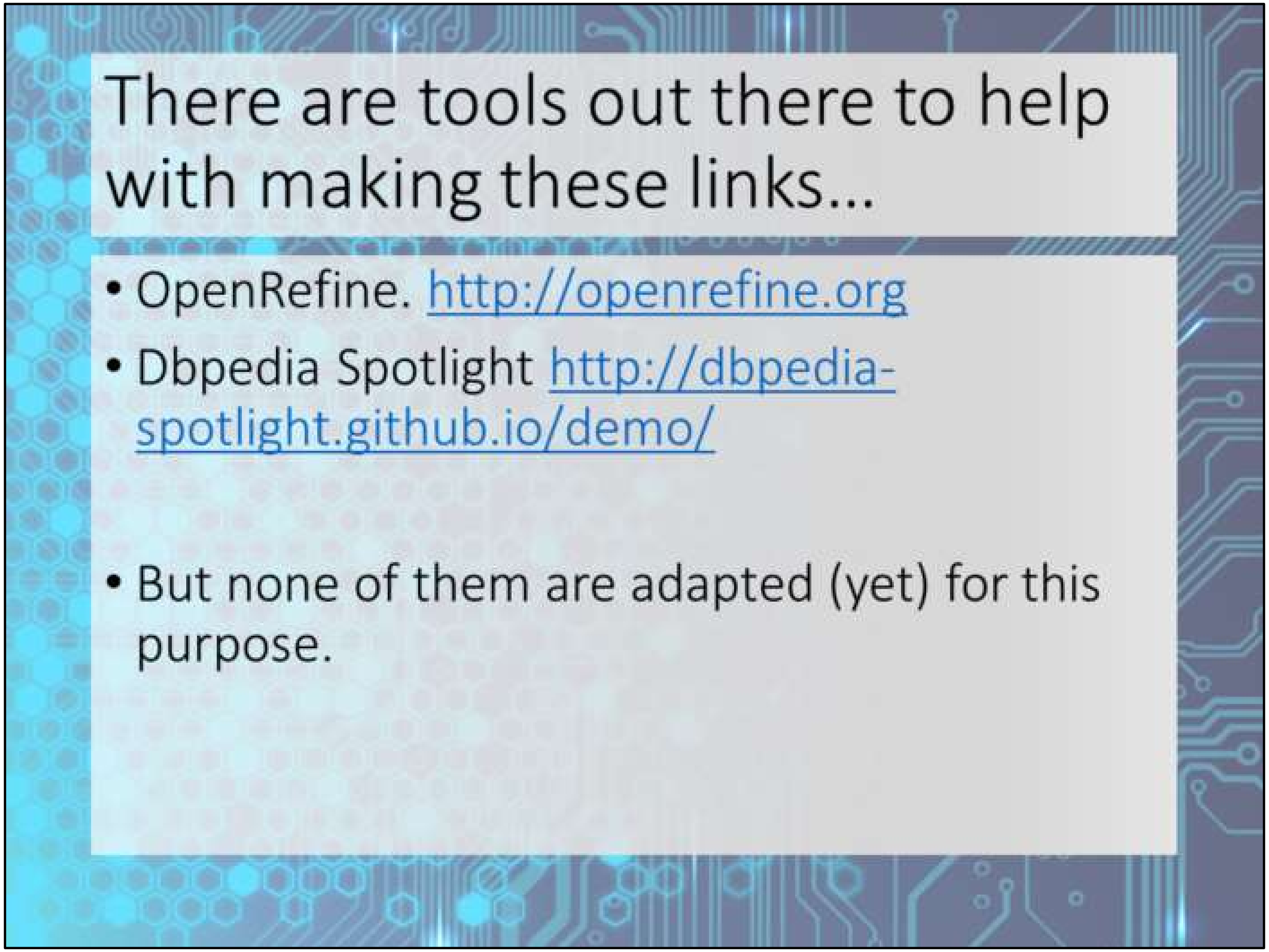
McCrae, John, Dennis Spohr & Philipp Cimiano. 2011. Linking Lexical Resources and Ontologies on the Semantic Web with lemon. *The Semantic Web: Research and Applications*, 245–259. http://pub.uni-bielefeld.de/luur/download?func=downloadFile&recordOId=2278403&fileOId=2526034.

Lemon is a proposed architecture for Lexicons in RDF and I used their model as a skeleton.



Future Work: Glosses

- Each entry could be linked to Wiktionary or Wikipedia pages in the analysis language for their referents.
- The linguist would have to do and verify this work, so this is not likely to happen until linguists see the value of it.
- These pages are linked to entries for the same referents in other languages.

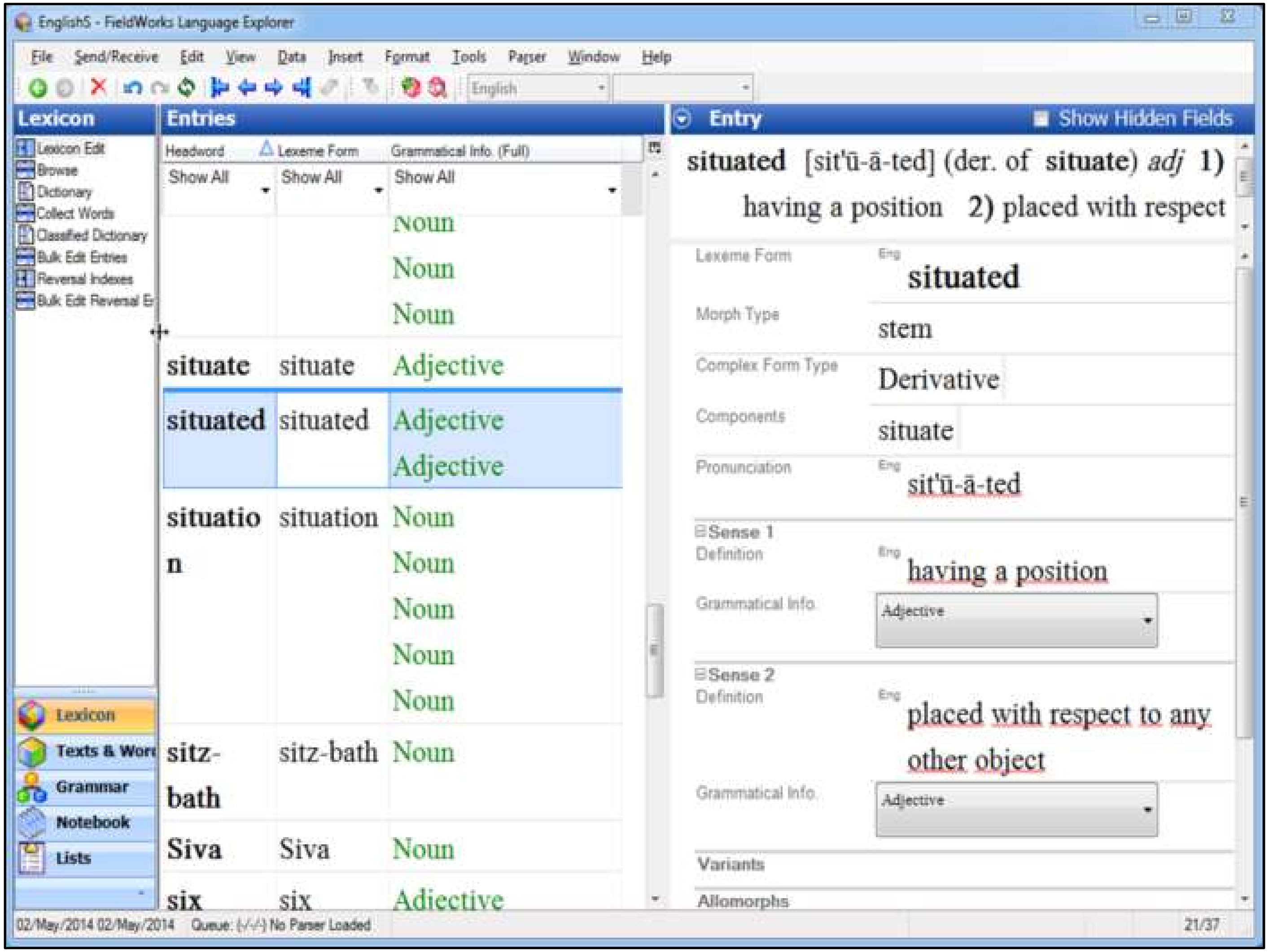


There are tools out there to help with making these links…

- OpenRefine. http://openrefine.org
- Dbpedia Spotlight http://dbpedia-spotlight.github.io/demo/

- But none of them are adapted (yet) for this purpose.

Dictionaries as Datapoints: Matthew Lee



Headword	Lexeme Form	Grammatical Info. (Full)
Show All	Show All	Show All
		Noun Noun Noun
situate	situate	Adjective
situated	situated	Adjective Adjective
situation	situation	Noun Noun Noun Noun Noun
sitz-bath	sitz-bath	Noun
Siva	Siva	Noun
six	six	Adjective

situated [sit'ū-ā-ted] (der. of **situate**) *adj* 1) having a position 2) placed with respect

Lexeme Form: Eng situated

Morph Type: stem

Complex Form Type: Derivative

Components: situate

Pronunciation: Eng sit'ū-ā-ted

Sense 1
Definition: Eng having a position
Grammatical Info.: Adjective

Sense 2
Definition: Eng placed with respect to any other object
Grammatical Info.: Adjective

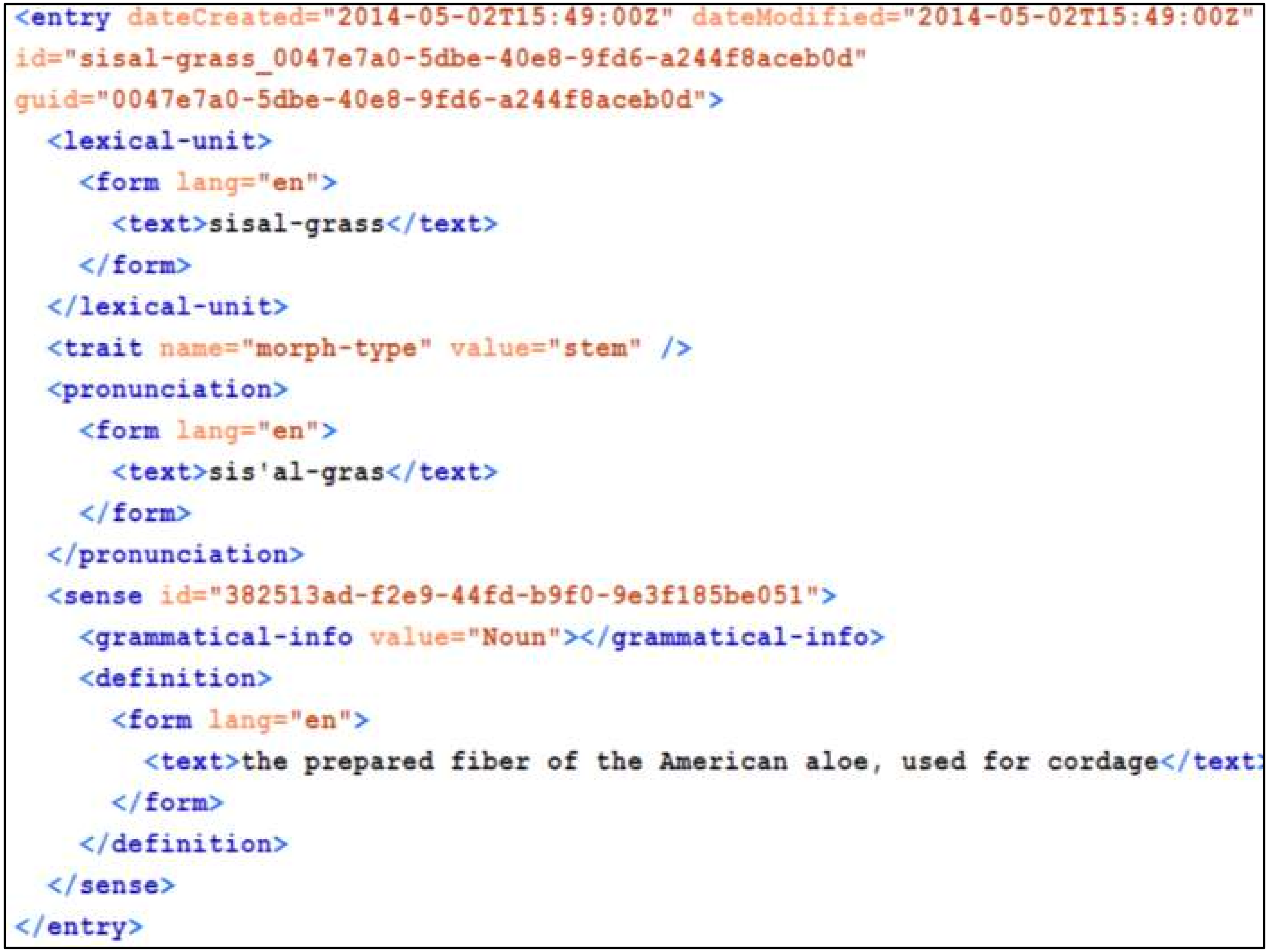
Variants

Allomorphs

02/May/2014 02/May/2014 Queue: (-/-/-) No Parser Loaded 21/37

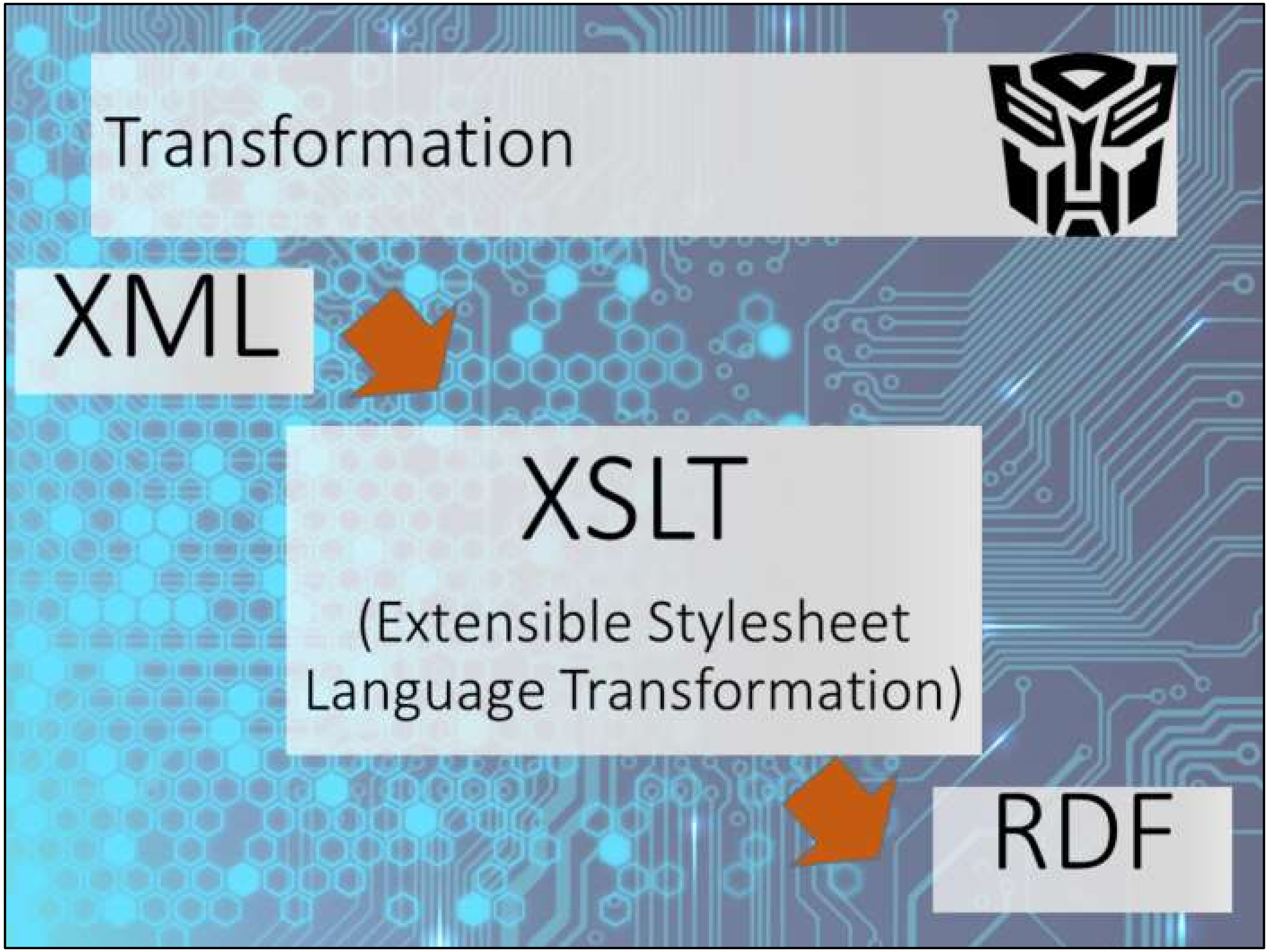
I take a Flex database, like this example database…and Export the parts that are important to me as LIFT XML.

Dictionaries as Datapoints: Matthew Lee



```
<entry dateCreated="2014-05-02T15:49:00Z" dateModified="2014-05-02T15:49:00Z"
id="sisal-grass_0047e7a0-5dbe-40e8-9fd6-a244f8aceb0d"
guid="0047e7a0-5dbe-40e8-9fd6-a244f8aceb0d">
  <lexical-unit>
    <form lang="en">
      <text>sisal-grass</text>
    </form>
  </lexical-unit>
  <trait name="morph-type" value="stem" />
  <pronunciation>
    <form lang="en">
      <text>sis'al-gras</text>
    </form>
  </pronunciation>
  <sense id="382513ad-f2e9-44fd-b9f0-9e3f185be051">
    <grammatical-info value="Noun"></grammatical-info>
    <definition>
      <form lang="en">
        <text>the prepared fiber of the American aloe, used for cordage</text
      </form>
    </definition>
  </sense>
</entry>
```

LIFT XML



Transformation

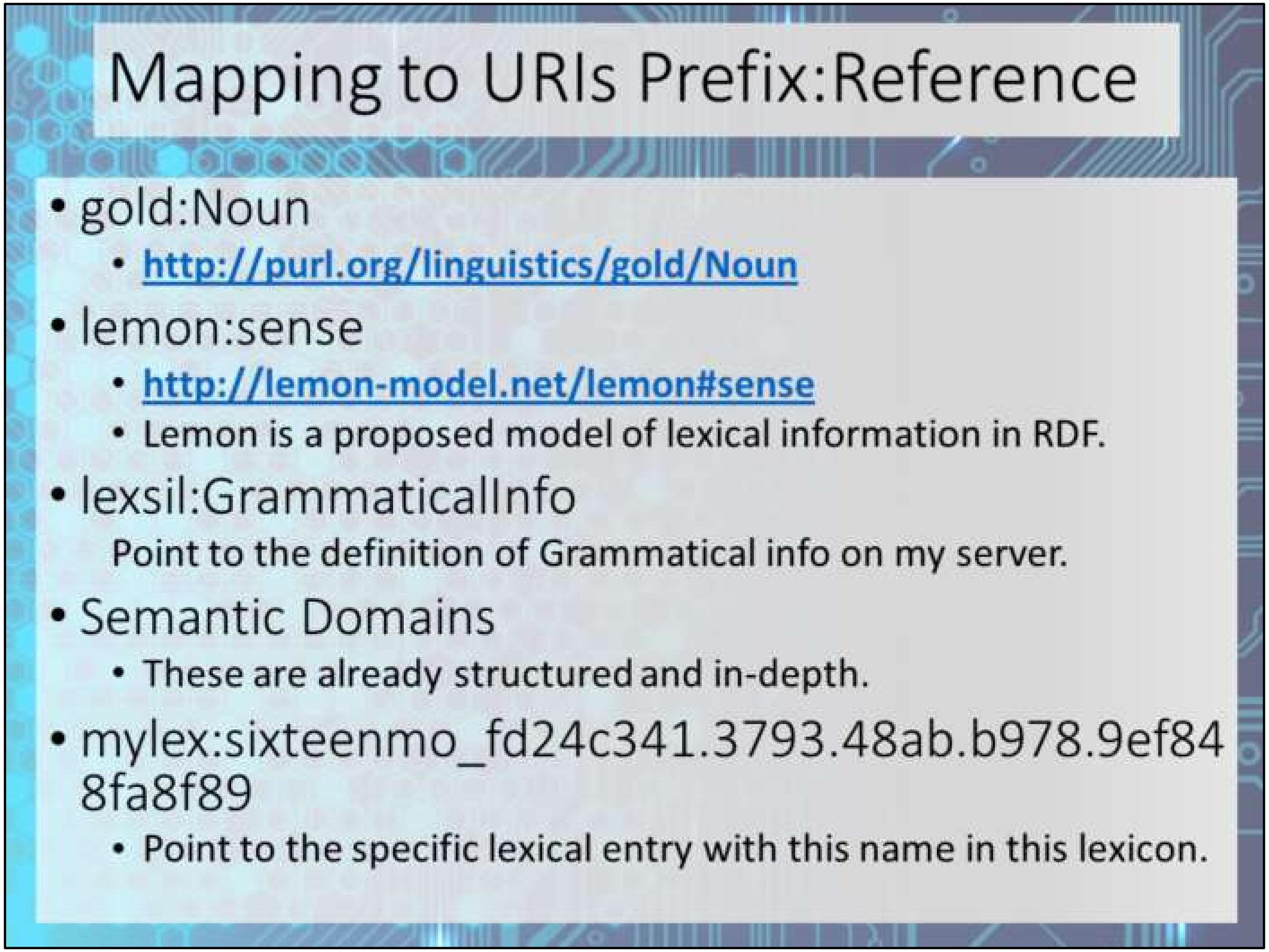
XML

XSLT

(Extensible Stylesheet Language Transformation)

RDF

This is where I spent most of my time on the project, tweaking the transformations to pull out the important structure and data.



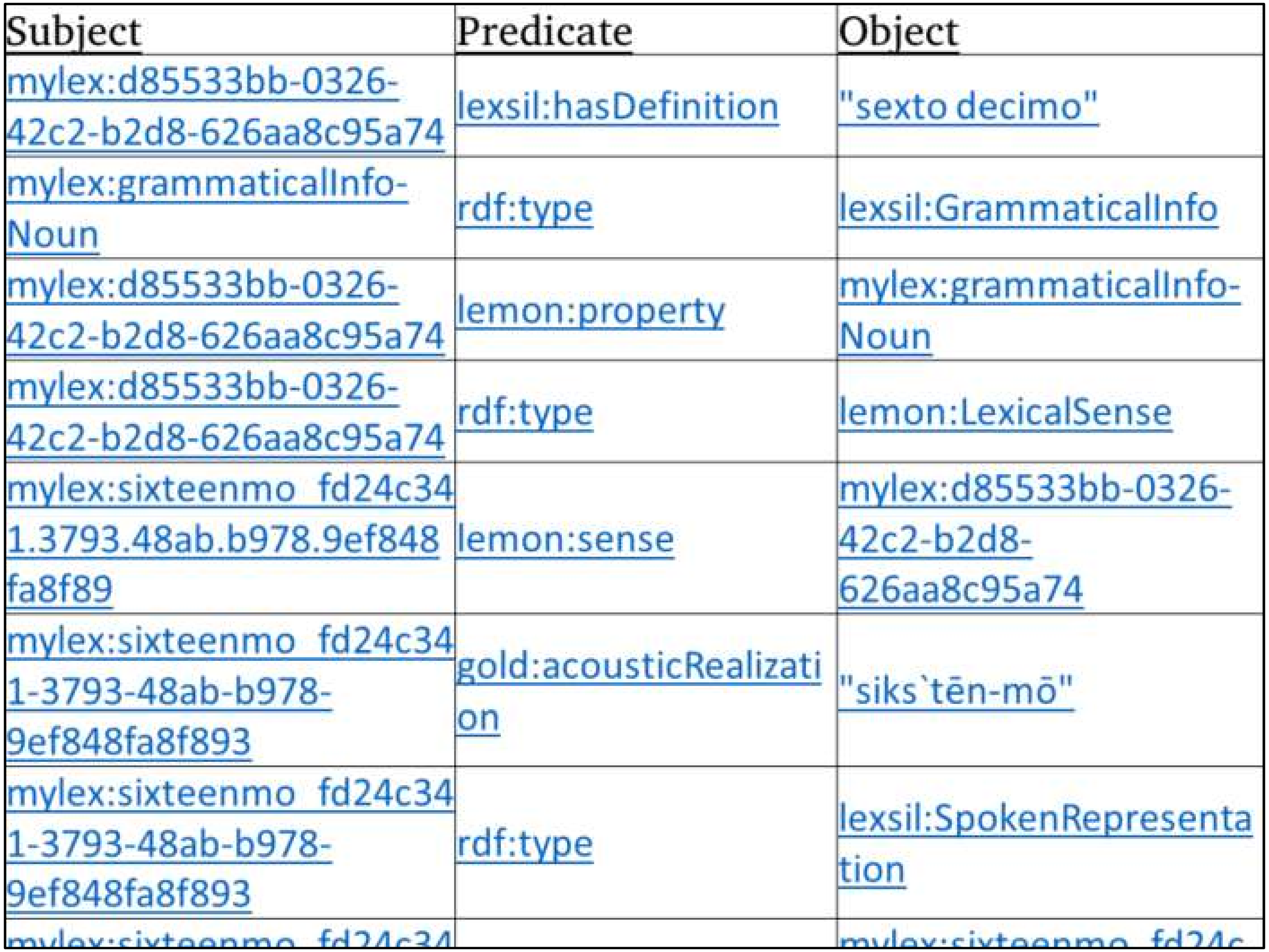
Mapping to URIs Prefix:Reference

- gold:Noun
  - http://purl.org/linguistics/gold/Noun
- lemon:sense
  - http://lemon-model.net/lemon#sense
  - Lemon is a proposed model of lexical information in RDF.
- lexsil:GrammaticalInfo

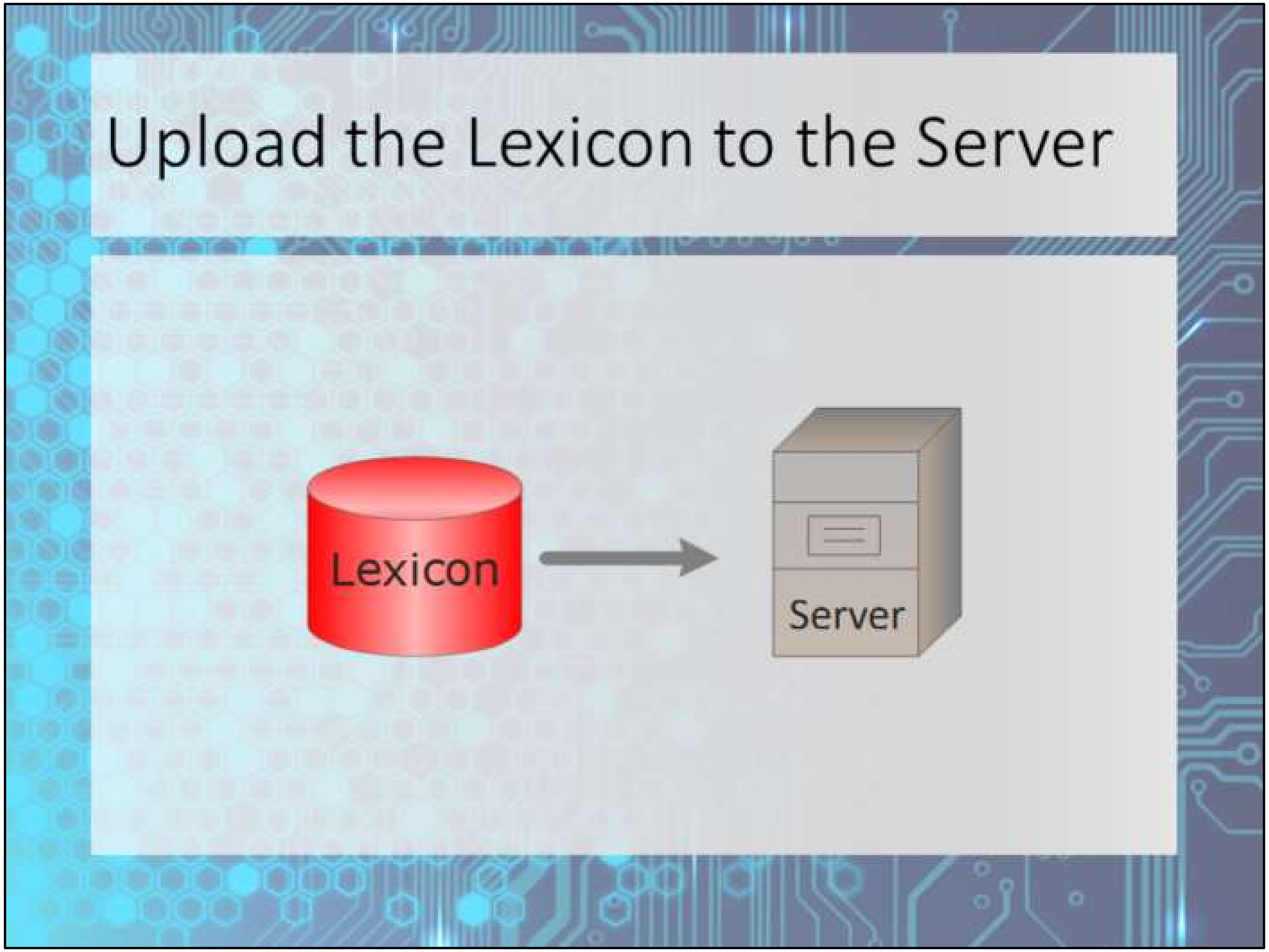
  Point to the definition of Grammatical info on my server.
- Semantic Domains
  - These are already structured and in-depth.
- mylex:sixteenmo_fd24c341.3793.48ab.b978.9ef848fa8f89
  - Point to the specific lexical entry with this name in this lexicon.

The Prefix, or part before the colon, is a shortcut pointing to an authority that catalogs such things, their catalog is called an Ontology.

Dictionaries as Datapoints: Matthew Lee

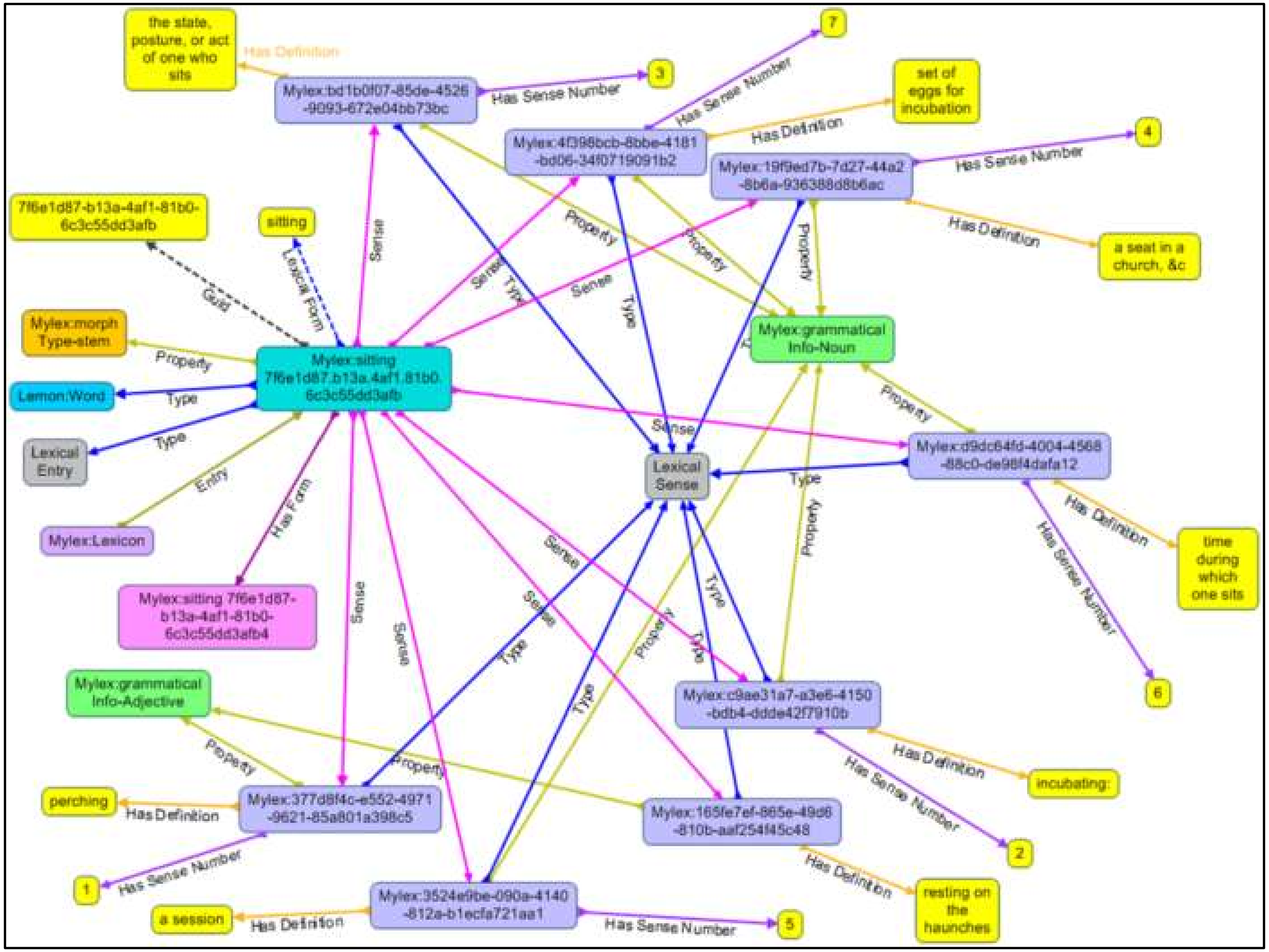


Subject	Predicate	Object
mylex:d85533bb-0326-42c2-b2d8-626aa8c95a74	lexsil:hasDefinition	"sexto decimo"
mylex:grammaticalInfo-Noun	rdf:type	lexsil:GrammaticalInfo
mylex:d85533bb-0326-42c2-b2d8-626aa8c95a74	lemon:property	mylex:grammaticalInfo-Noun
mylex:d85533bb-0326-42c2-b2d8-626aa8c95a74	rdf:type	lemon:LexicalSense
mylex:sixteenmo_fd24c341.3793.48ab.b978.9ef848fa8f89	lemon:sense	mylex:d85533bb-0326-42c2-b2d8-626aa8c95a74
mylex:sixteenmo_fd24c341-3793-48ab-b978-9ef848fa8f893	gold:acousticRealization	"siks`tēn-mō"
mylex:sixteenmo_fd24c341-3793-48ab-b978-9ef848fa8f893	rdf:type	lexsil:SpokenRepresentation
mylex:sixteenmo_fd24c34		mylex:sixteenmo_fd24c



Upload the Lexicon to the Server

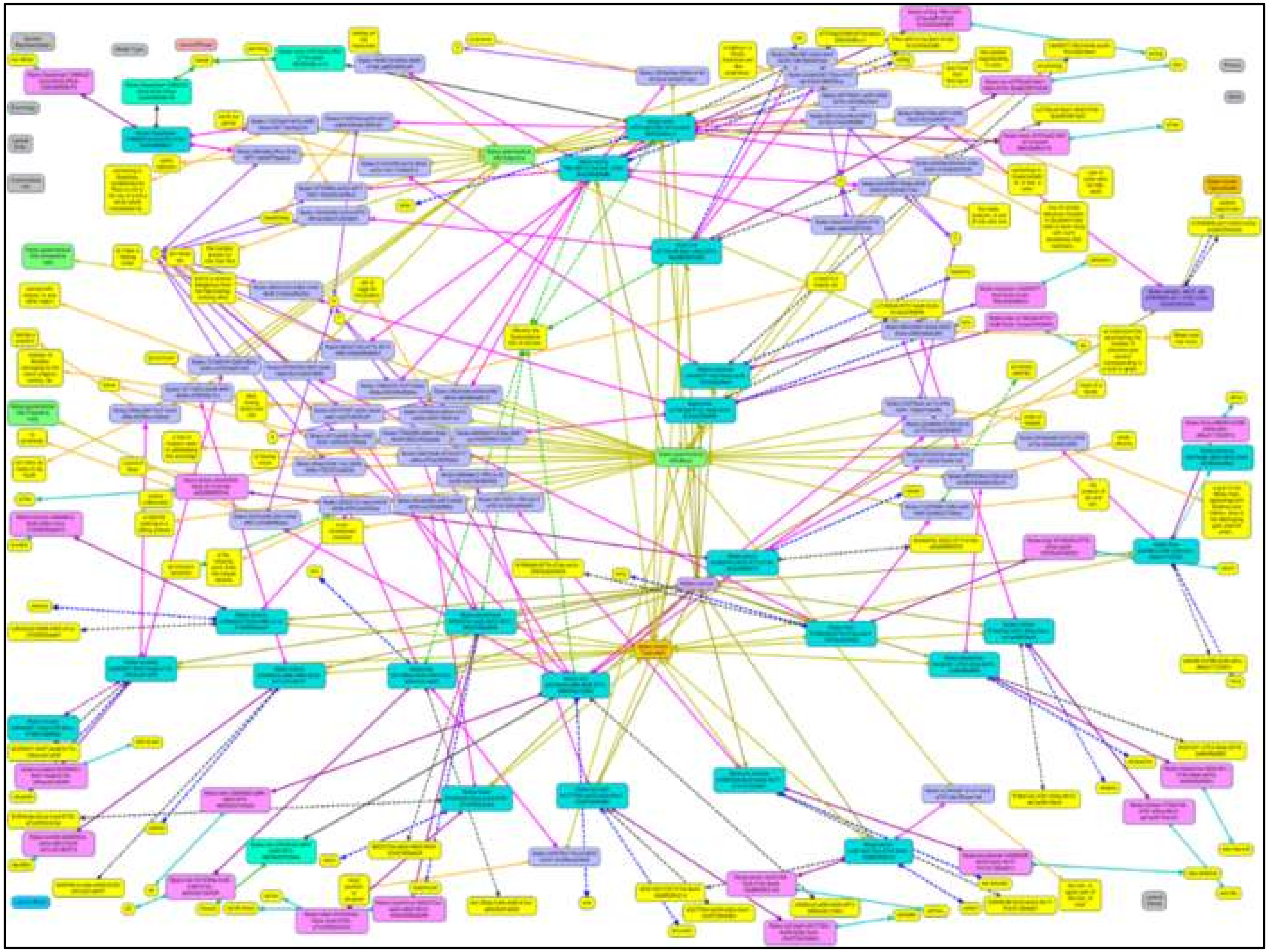
Dictionaries as Datapoints: Matthew Lee



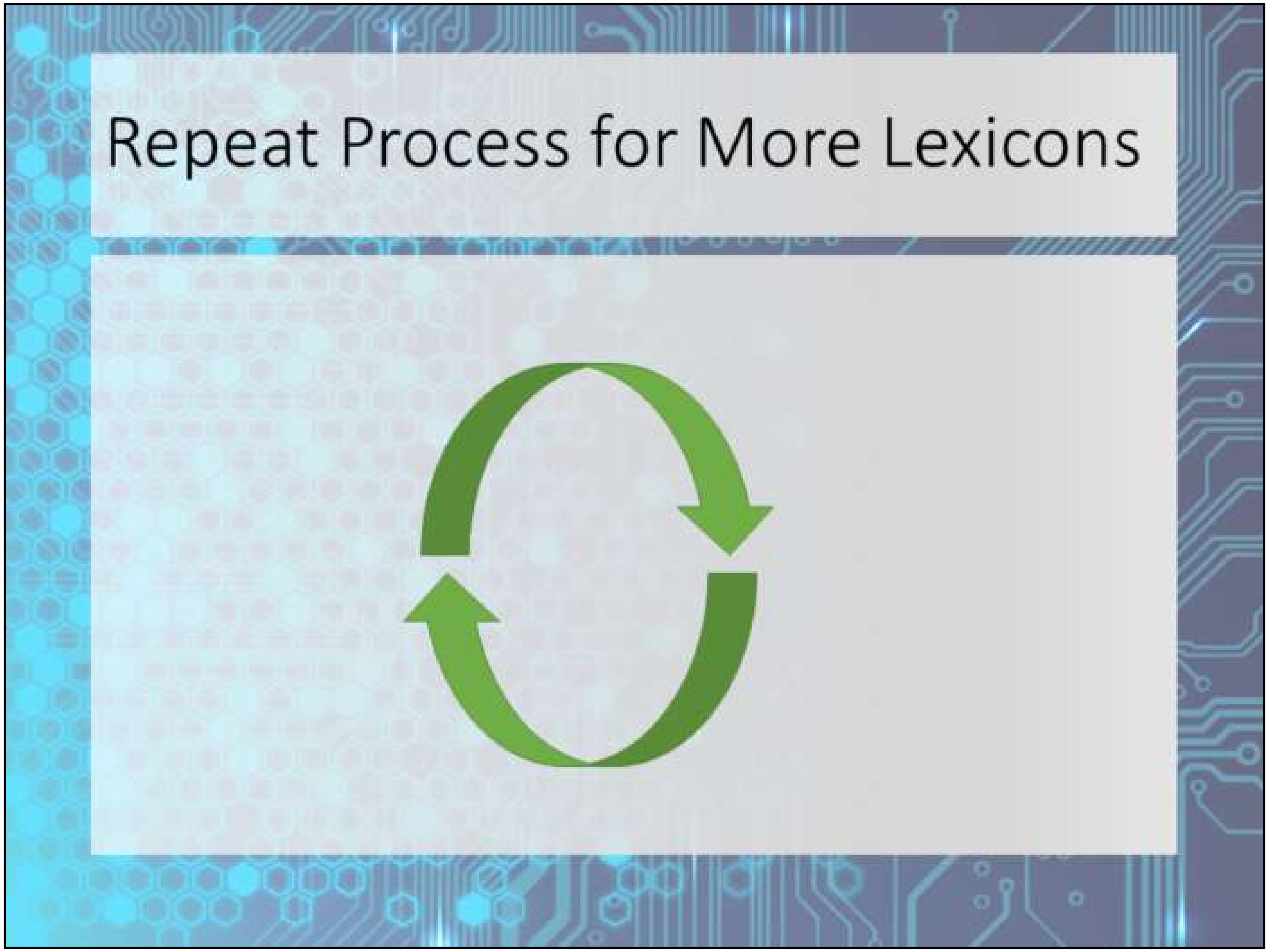
the state, posture, or act of one who sits
Has Definition
Mylex:bd1b0f07-85de-4526 -9093-672e04bb73bc
Has Sense Number
3
7
Has Sense Number
set of eggs for incubation
Mylex:4f398bcb-8bbe-4181 -bd06-34f0719091b2
Has Definition
Mylex:19f9ed7b-7d27-44a2 -8b6a-936388d8b6ac
Has Sense Number
4
7f6e1d87-b13a-4af1-81b0-6c3c55dd3afb
sitting
Has Definition
a seat in a church, &c
Guid
Lexical Form
Sense
Sense
Type
Property
Sense
Type
Property
Property
Mylex:morph Type-stem
Property
Mylex:sitting 7f6e1d87.b13a.4af1.81b0. 6c3c55dd3afb
Mylex:grammatical Info-Noun
Lemon:Word
Type
Type
Lexical Entry
Sense
Property
Mylex:d9dc64fd-4004-4568 -88c0-de98f4dafa12
Entry
Lexical Sense
Type
Mylex:Lexicon
Has Form
Property
Has Definition
Has Sense Number
time during which one sits
Sense
Mylex:sitting 7f6e1d87-b13a-4af1-81b0-6c3c55dd3afb4
Sense
Sense
Sense
Type
Type
Property
Type
6
Mylex:grammatical Info-Adjective
Mylex:c9ae31a7-a3e6-4150 -bdb4-ddde42f7910b
Type
Property
Property
Has Definition
incubating:
Has Sense Number
perching
Has Definition
Mylex:377d8f4c-e552-4971 -9621-85a801a398c5
Mylex:165fe7ef-865e-49d6 -810b-aaf254f45c48
2
Has Sense Number
1
Has Definition
resting on the haunches
a session
Has Definition
Mylex:3524e9be-090a-4140 -812a-b1ecfa721aa1
Has Sense Number
5

This is a single entry from the example database from before.

Dictionaries as Datapoints: Matthew Lee

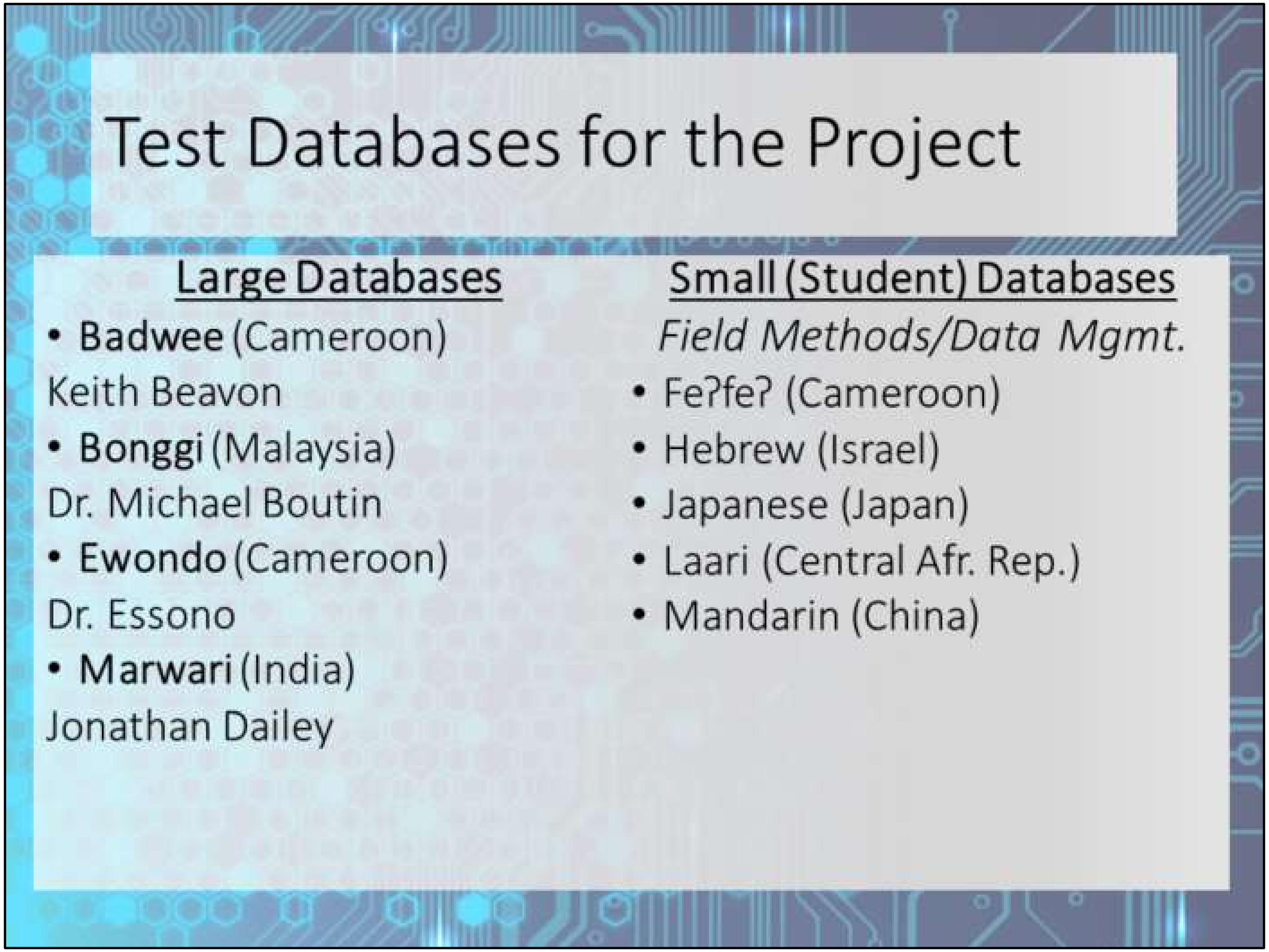


This is all of the data from the dictionary page before. Once the relationships are defined, we have 884 Facts.



Repeat Process for More Lexicons

This is now an automated conversion…so I just have to point the process to an exported lexicon and push play.



Test Databases for the Project

Large Databases

- **Badwee** (Cameroon)

Keith Beavon

- **Bonggi** (Malaysia)

Dr. Michael Boutin

- **Ewondo** (Cameroon)

Dr. Essono

- **Marwari** (India)

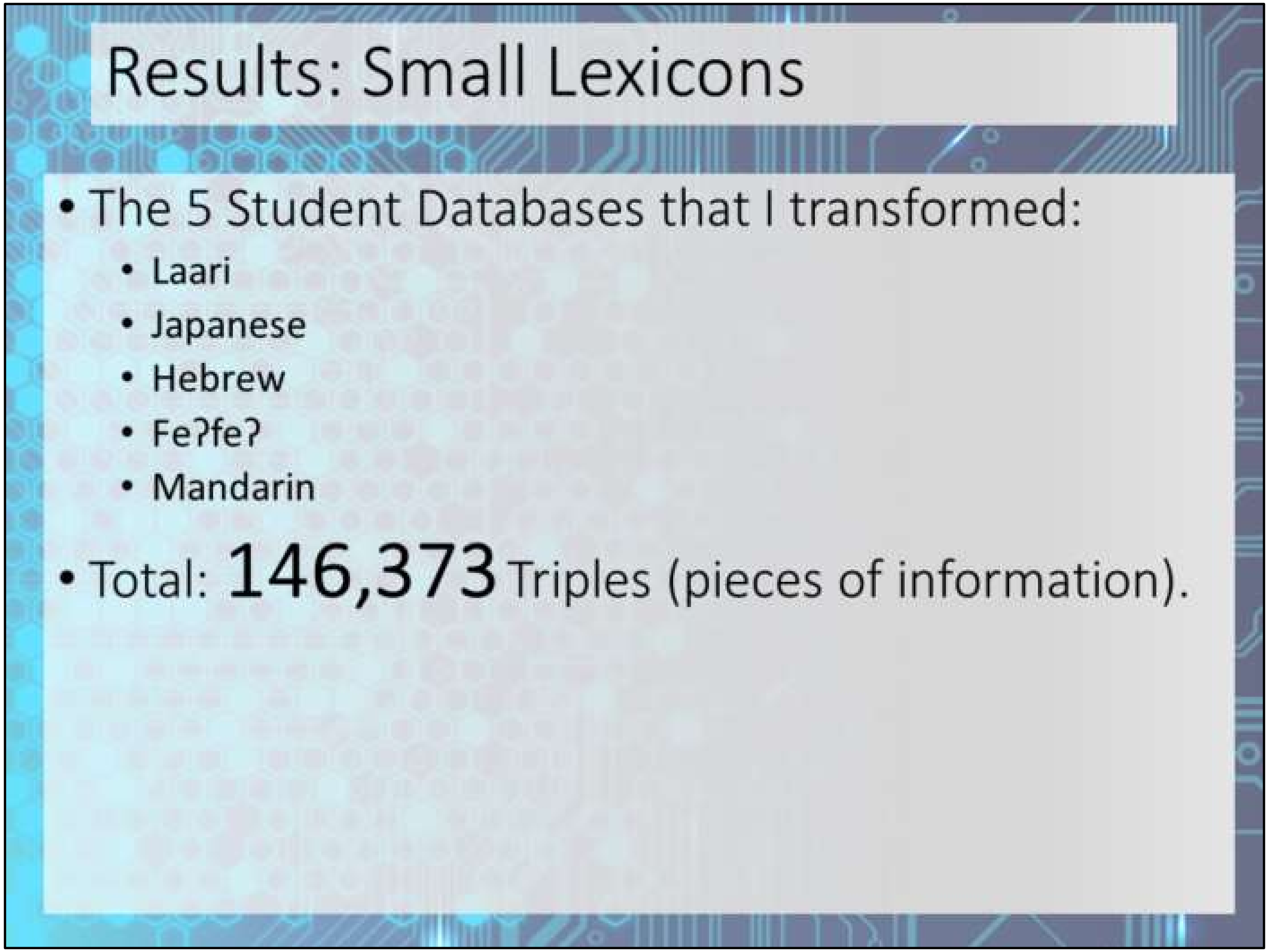
Jonathan Dailey

Small (Student) Databases

Field Methods/Data Mgmt.

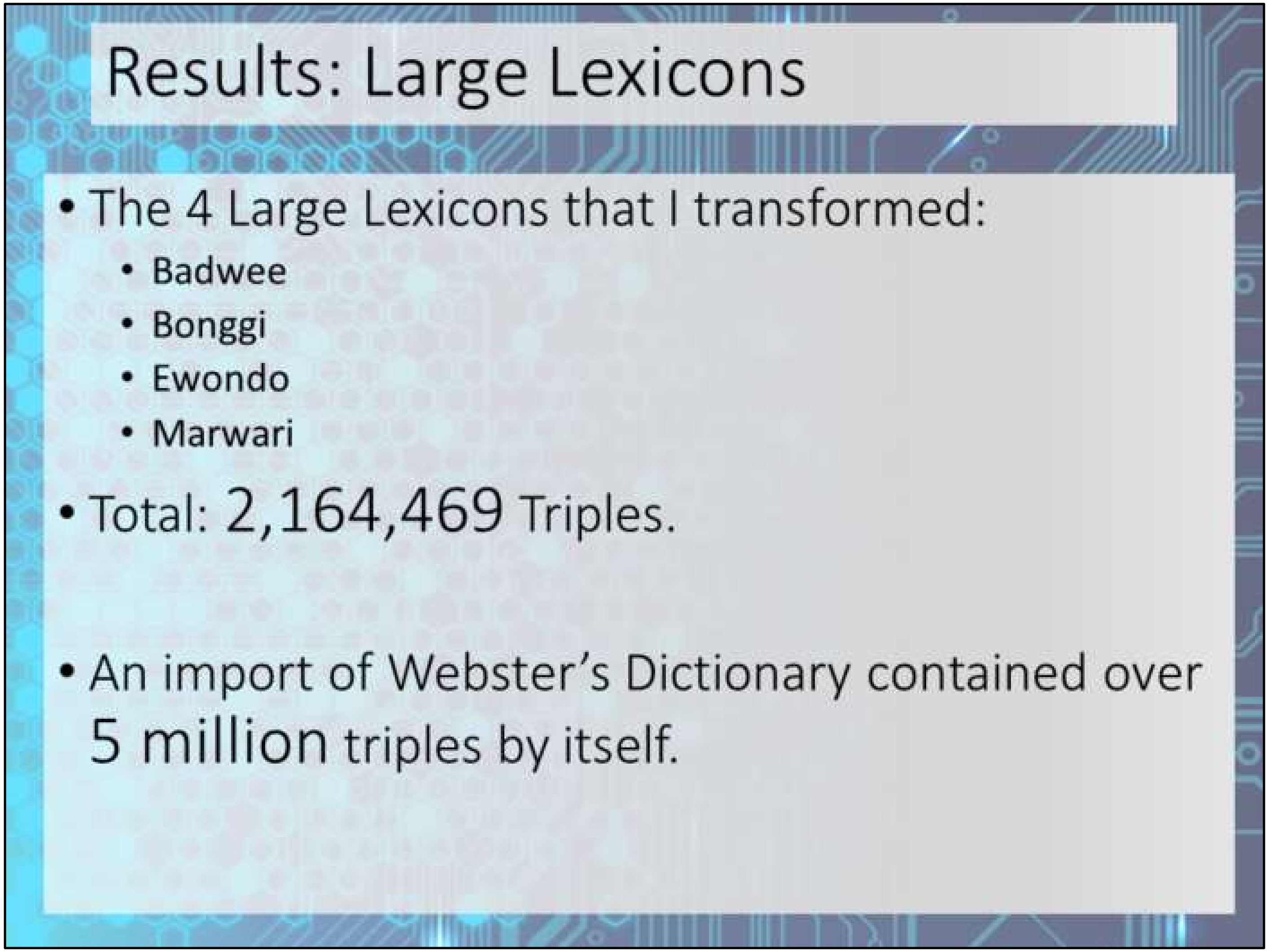
- Fe?fe? (Cameroon)
- Hebrew (Israel)
- Japanese (Japan)
- Laari (Central Afr. Rep.)
- Mandarin (China)

Results:



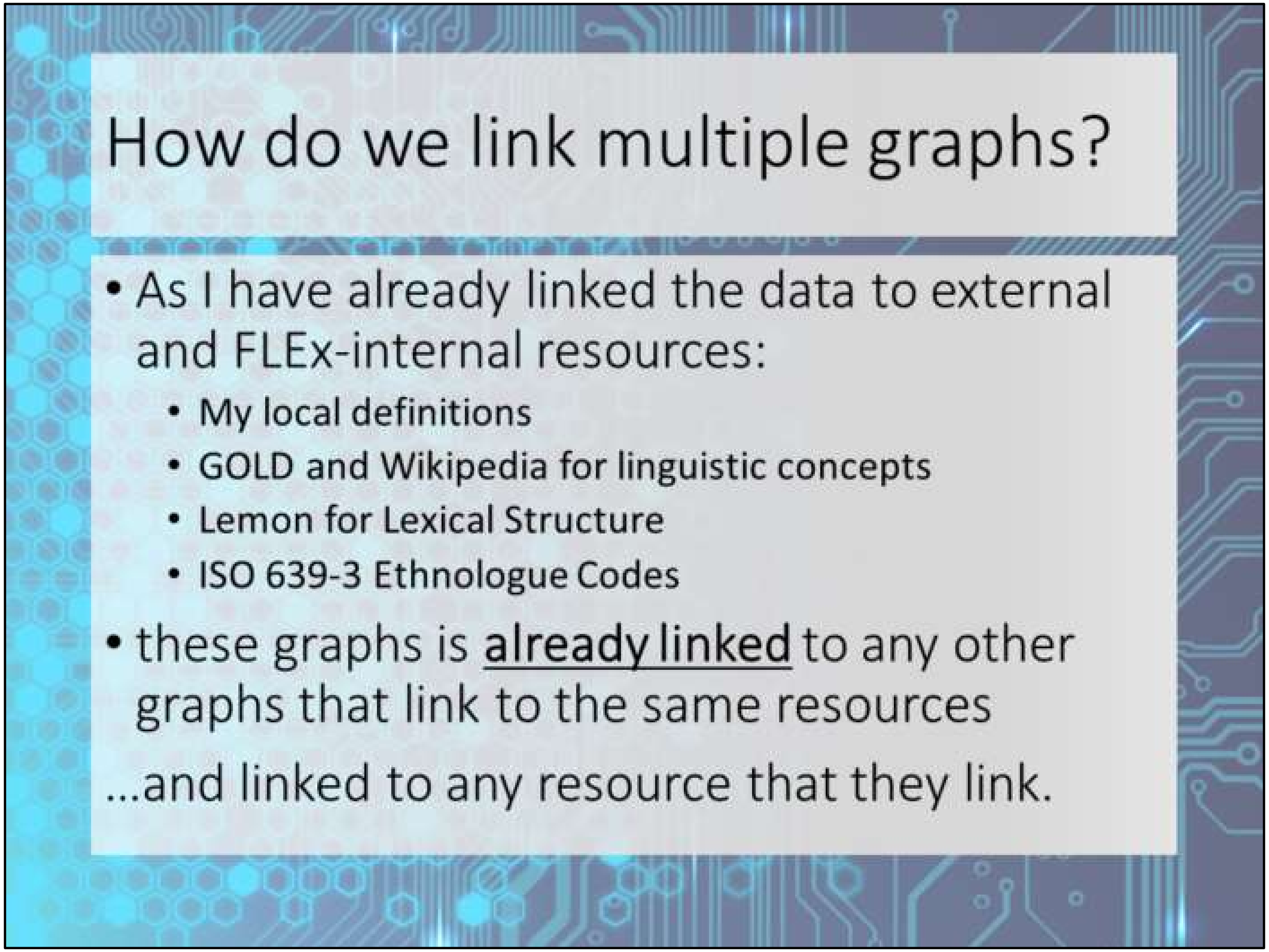
Results: Small Lexicons

- The 5 Student Databases that I transformed:
  - Laari
  - Japanese
  - Hebrew
  - Feʔfeʔ
  - Mandarin
- Total: 146,373 Triples (pieces of information).



Results: Large Lexicons

- The 4 Large Lexicons that I transformed:
  - Badwee
  - Bonggi
  - Ewondo
  - Marwari
- Total: 2,164,469 Triples.
- An import of Webster's Dictionary contained over 5 million triples by itself.



How do we link multiple graphs?

- As I have already linked the data to external and FLEx-internal resources:
  - My local definitions
  - GOLD and Wikipedia for linguistic concepts
  - Lemon for Lexical Structure
  - ISO 639-3 Ethnologue Codes
- these graphs is already linked to any other graphs that link to the same resources

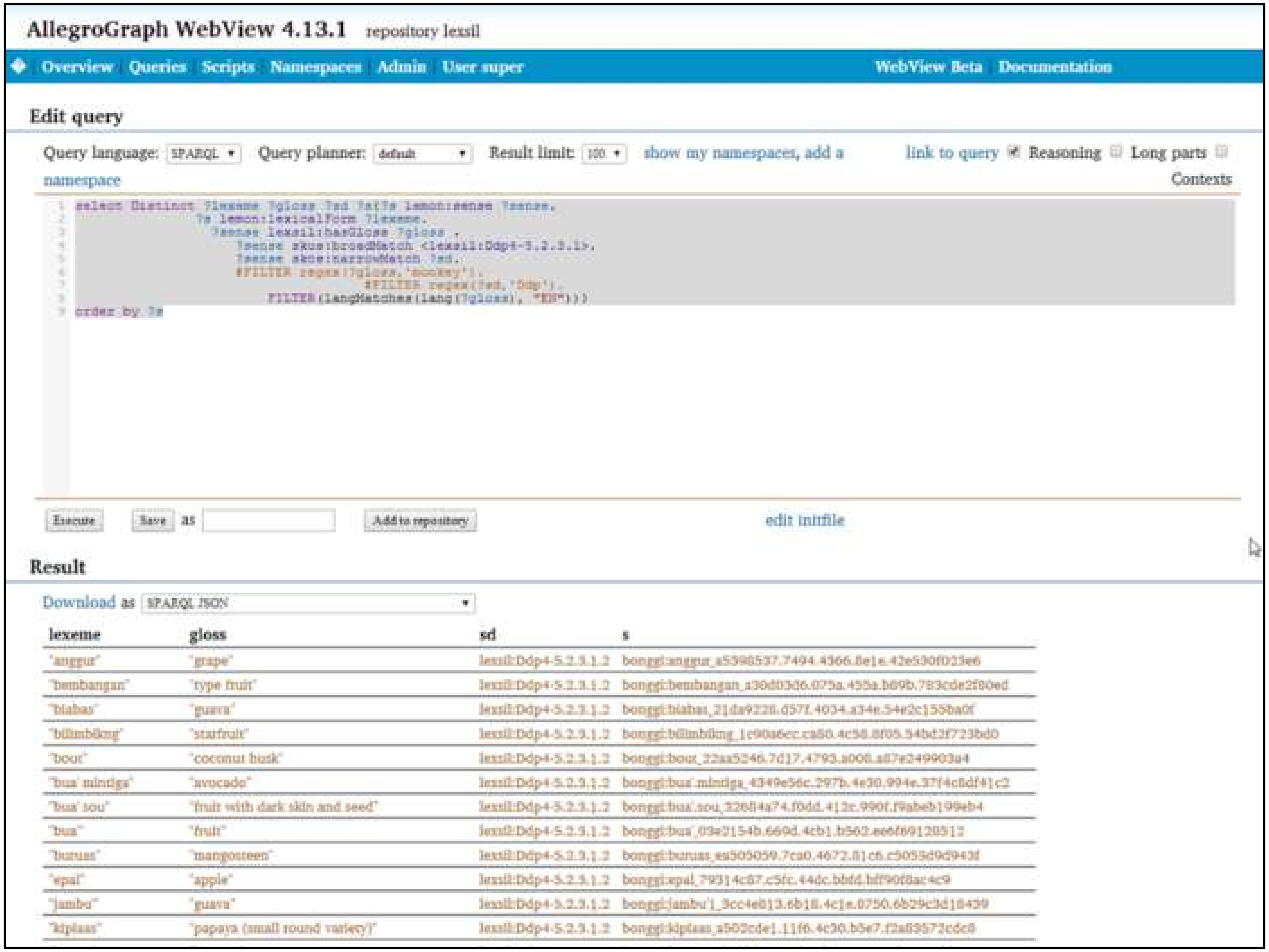
…and linked to any resource that they link.

Ontology

Queries and Questions

Now that the data is linked, we can start to ask questions…

Dictionaries as Datapoints: Matthew Lee



AllegroGraph WebView 4.13.1 repository lexsil
Overview Queries Scripts Namespaces Admin User super
WebView Beta Documentation

Edit query

Query language: SPARQL ▾ Query planner: default ▾ Result limit: 100 ▾ show my namespaces, add a namespace link to query ☑ Reasoning □ Long parts □ Contexts

```
select Distinct ?lexeme ?gloss ?sd ?s{?s lemon:sense ?sense.
                 ?s lemon:lexicalForm ?lexeme.
                   ?sense lexsil:hasGloss ?gloss .
                      ?sense skos:broadMatch <lexsil:Ddp4-5.2.3.1>.
                      ?sense skos:narrowMatch ?sd.
                      #FILTER regex(?gloss, 'monkey').
                                       #FILTER regex(?sd,'Ddp').
                          FILTER(langMatches(lang(?gloss), "EN")))
order by ?s
```

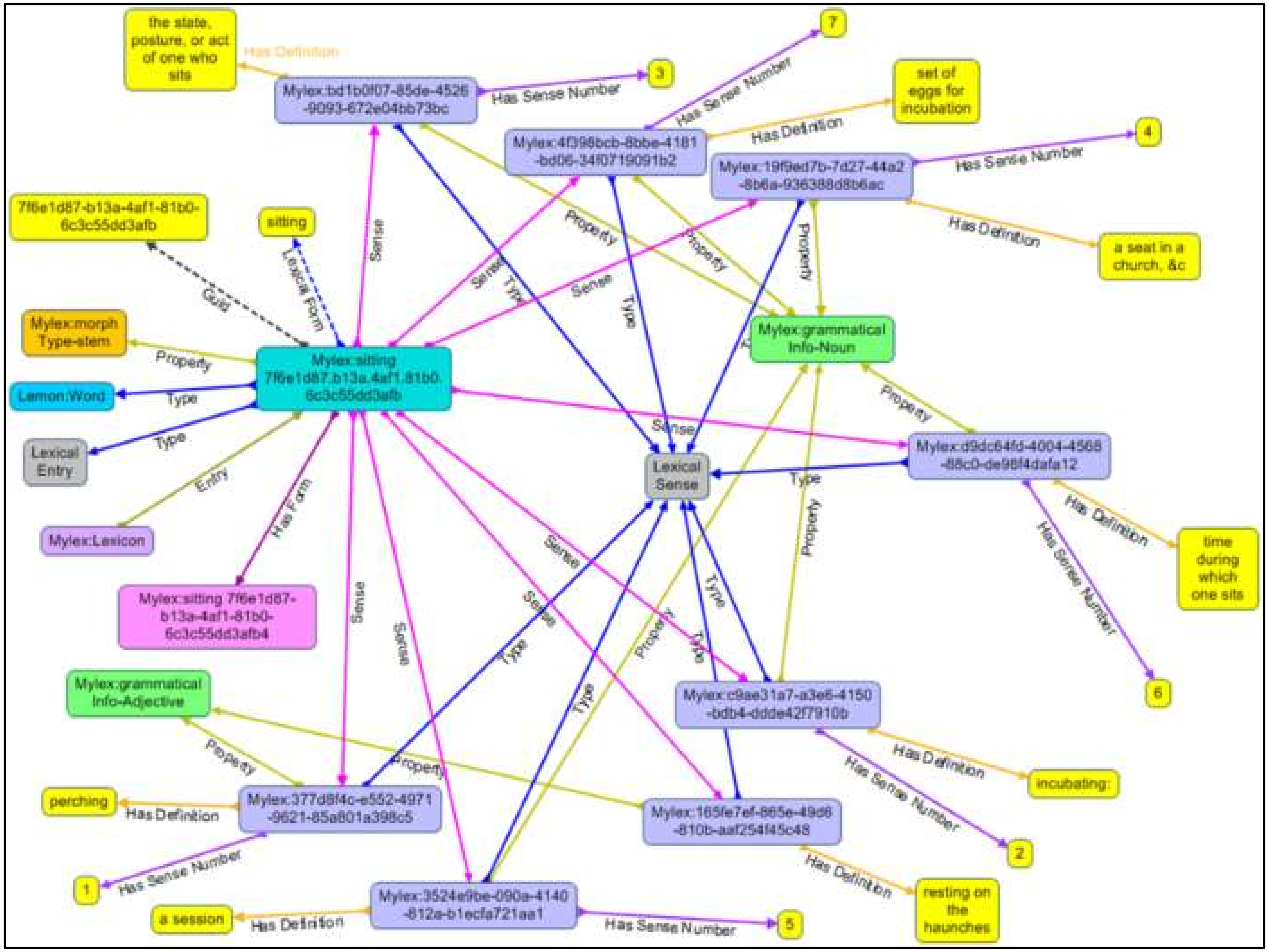
Result

Download as SPARQL JSON ▾

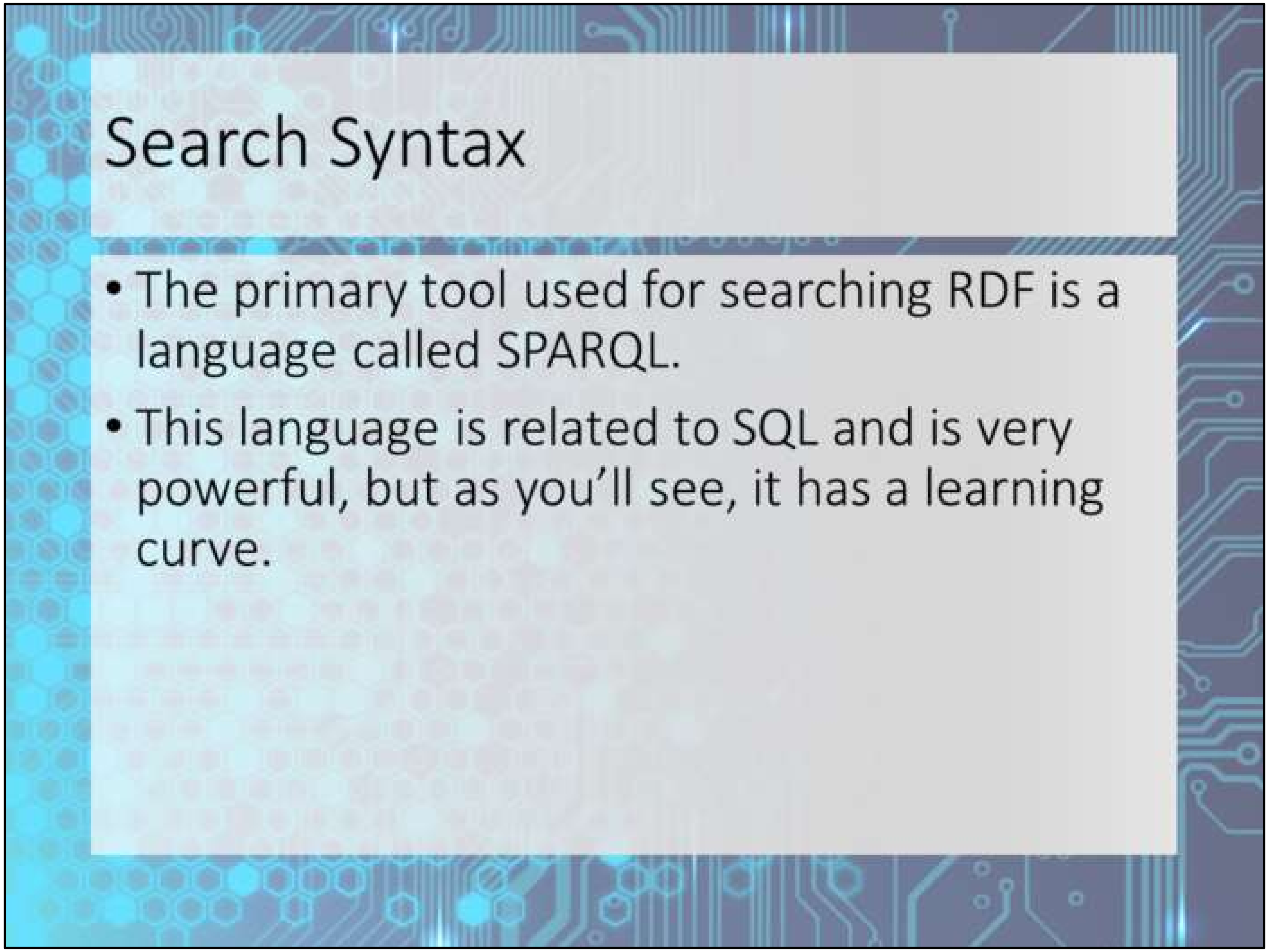
lexeme	gloss	sd	s
"anggur"	"grape"	lexsil:Ddp4-5.2.3.1.2	bonggi:anggur_a5398537.7494.4366.8e1e.42e530f023e6
"bembangan"	"type fruit"	lexsil:Ddp4-5.2.3.1.2	bonggi:bembangan_a30d03d6.075a.455a.b69b.783cde2f80ed
"biabas"	"guava"	lexsil:Ddp4-5.2.3.1.2	bonggi:biabas_21da9228.d57f.4034.a34e.54e2c155ba0f
"bilimbing"	"starfruit"	lexsil:Ddp4-5.2.3.1.2	bonggi:bilimbing_1c90a6cc.ca80.4c58.8f05.54bd3f723bd0
"bour"	"coconut husk"	lexsil:Ddp4-5.2.3.1.2	bonggi:bour_22aa5246.7d17.4793.a008.a87e249903a4
"bua' mintiga"	"avocado"	lexsil:Ddp4-5.2.3.1.2	bonggi:bua'mintiga_4349e56c.297b.4e30.994e.37f4c8df41c2
"bua' sou"	"fruit with dark skin and seed"	lexsil:Ddp4-5.2.3.1.2	bonggi:bua'sou_32684a74.f0dd.412c.990f.f9abeb199eb4
"bua"	"fruit"	lexsil:Ddp4-5.2.3.1.2	bonggi:bua_03e2154b.669d.4cb1.b562.ee6f69128512
"buruas"	"mangosteen"	lexsil:Ddp4-5.2.3.1.2	bonggi:buruas_ea505059.7ca0.4672.81c6.c5053d9d943f
"epal"	"apple"	lexsil:Ddp4-5.2.3.1.2	bonggi:epal_79314c87.c5fc.44dc.bbfd.bff90f8ac4c9
"jambu"	"guava"	lexsil:Ddp4-5.2.3.1.2	bonggi:jambu'1_3cc4e813.6b18.4c1e.0750.6b29c3d18439
"kiplaas"	"papaya (small round variety)"	lexsil:Ddp4-5.2.3.1.2	bonggi:kiplaas_a502cde1.11f6.4c30.b5e7.f2a83572cdc8

The primary database I used, Allegrograph, has two interfaces: this one for searching, and [click]

Dictionaries as Datapoints: Matthew Lee

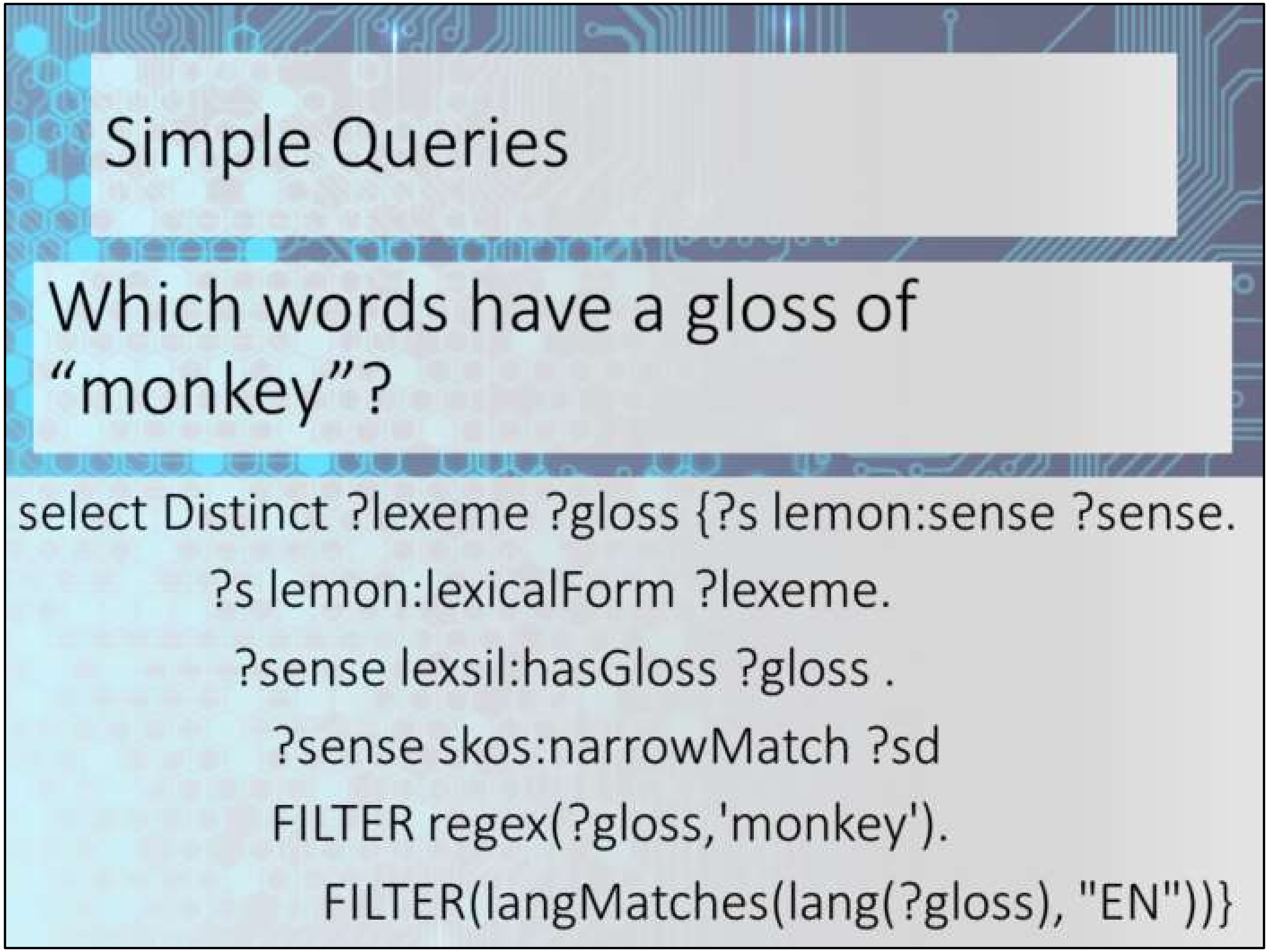

the state, posture, or act of one who sits
Has Definition
Mylex:bd1b0f07-85de-4526-9093-672e04bb73bc
Has Sense Number
3
7
Has Sense Number
set of eggs for incubation
Has Definition
Mylex:4f398bcb-8bbe-4181-bd06-34f0719091b2
Mylex:19f9ed7b-7d27-44a2-8b6a-936388d8b6ac
Has Sense Number
4
Has Definition
a seat in a church, &c
7f6e1d87-b13a-4af1-81b0-6c3c55dd3afb
sitting
Guid
Lexical Form
Sense
Property
Type
Mylex:morph Type-stem
Property
Mylex:sitting 7f6e1d87.b13a.4af1.81b0.6c3c55dd3afb
Mylex:grammatical Info-Noun
Lemon:Word
Type
Lexical Entry
Type
Entry
Mylex:Lexicon
Has Form
Lexical Sense
Mylex:d9dc64fd-4004-4568-88c0-de98f4dafa12
Has Definition
time during which one sits
Has Sense Number
6
Mylex:sitting 7f6e1d87-b13a-4af1-81b0-6c3c55dd3afb4
Mylex:grammatical Info-Adjective
Mylex:c9ae31a7-a3e6-4150-bdb4-ddde42f7910b
Has Definition
incubating:
Has Sense Number
2
perching
Has Definition
Mylex:377d8f4c-e552-4971-9621-85a801a398c5
Mylex:165fe7ef-865e-49d6-810b-aaf254f45c48
Has Definition
resting on the haunches
1
Has Sense Number
a session
Has Definition
Mylex:3524e9be-090a-4140-812a-b1ecfa721aa1
Has Sense Number
5

this one for Browsing



Search Syntax

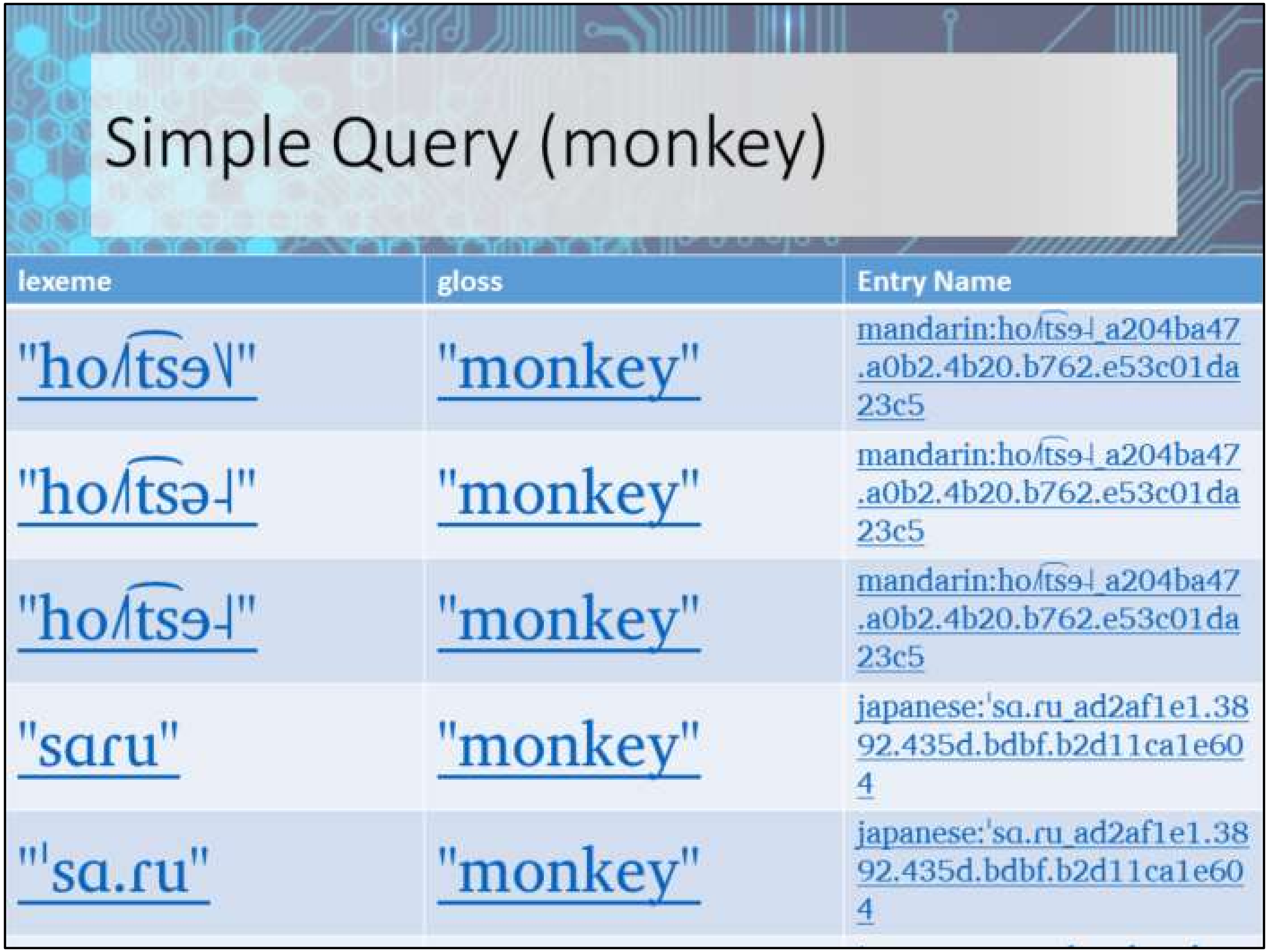
- The primary tool used for searching RDF is a language called SPARQL.
- This language is related to SQL and is very powerful, but as you'll see, it has a learning curve.



Simple Queries

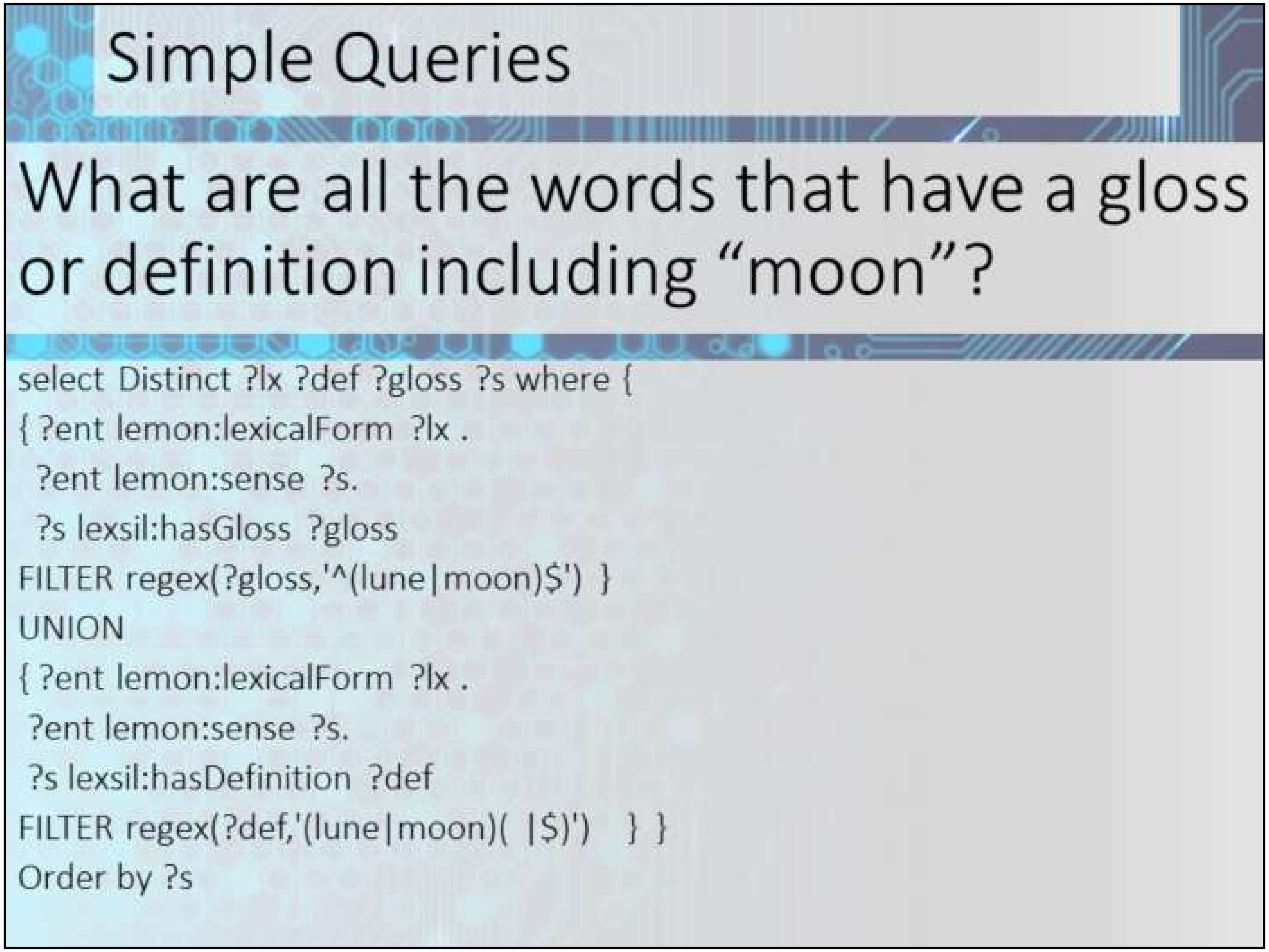
Which words have a gloss of "monkey"?

```
select Distinct ?lexeme ?gloss {?s lemon:sense ?sense.
           ?s lemon:lexicalForm ?lexeme.
            ?sense lexsil:hasGloss ?gloss .
              ?sense skos:narrowMatch ?sd
              FILTER regex(?gloss,'monkey').
                 FILTER(langMatches(lang(?gloss), "EN"))}
```



Simple Query (monkey)

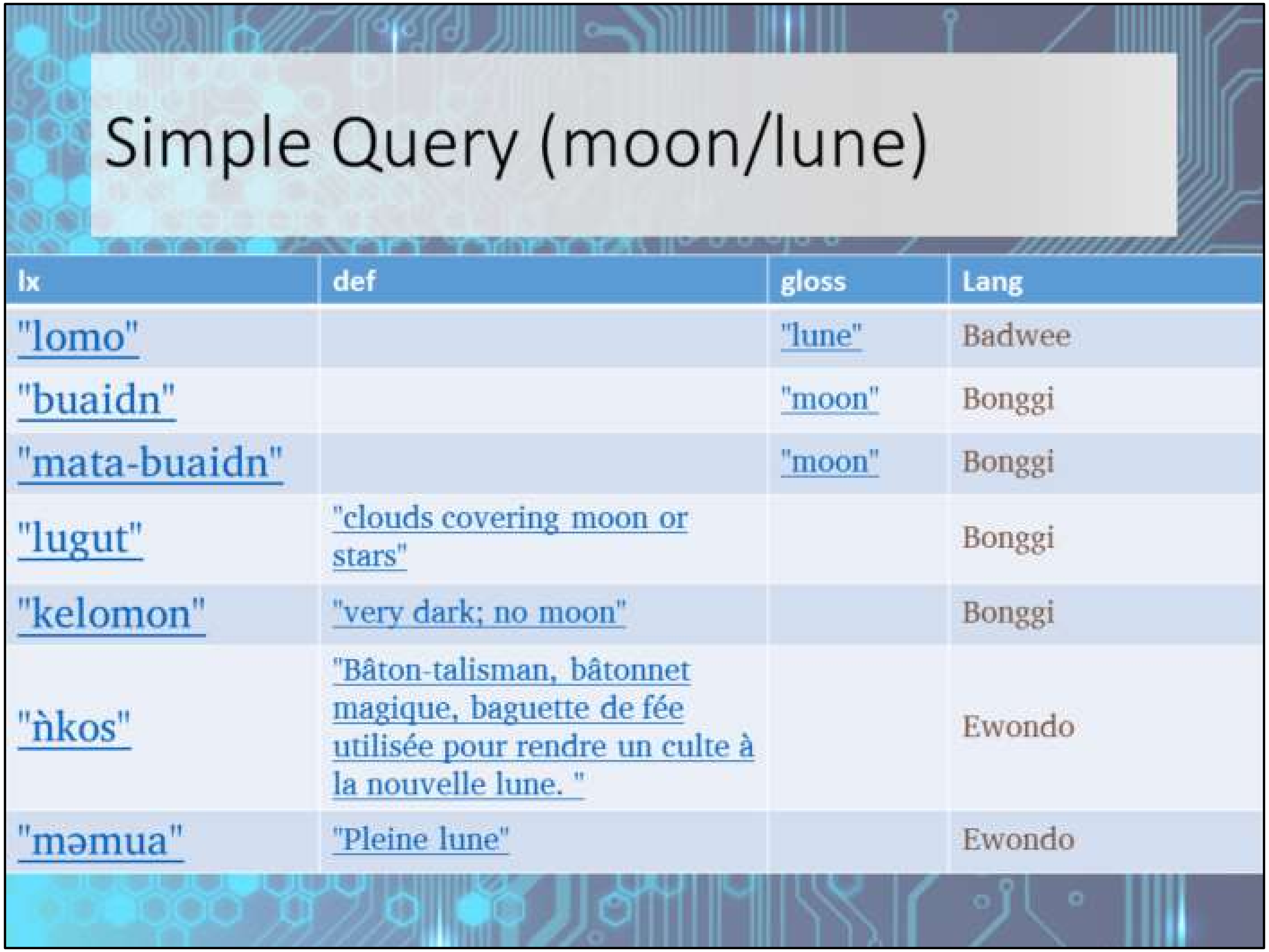
lexeme	gloss	Entry Name
"ho˧˥t͡sə˥˩"	"monkey"	mandarin:ho˧˥t͡sə˨_a204ba47.a0b2.4b20.b762.e53c01da23c5
"ho˧˥t͡sə˧"	"monkey"	mandarin:ho˧˥t͡sə˨_a204ba47.a0b2.4b20.b762.e53c01da23c5
"ho˧˥t͡sə˧"	"monkey"	mandarin:ho˧˥t͡sə˨_a204ba47.a0b2.4b20.b762.e53c01da23c5
"saɾu"	"monkey"	japanese:ˈsɑ.ɾu_ad2af1e1.3892.435d.bdbf.b2d11ca1e604
"ˈsɑ.ɾu"	"monkey"	japanese:ˈsɑ.ɾu_ad2af1e1.3892.435d.bdbf.b2d11ca1e604



Simple Queries

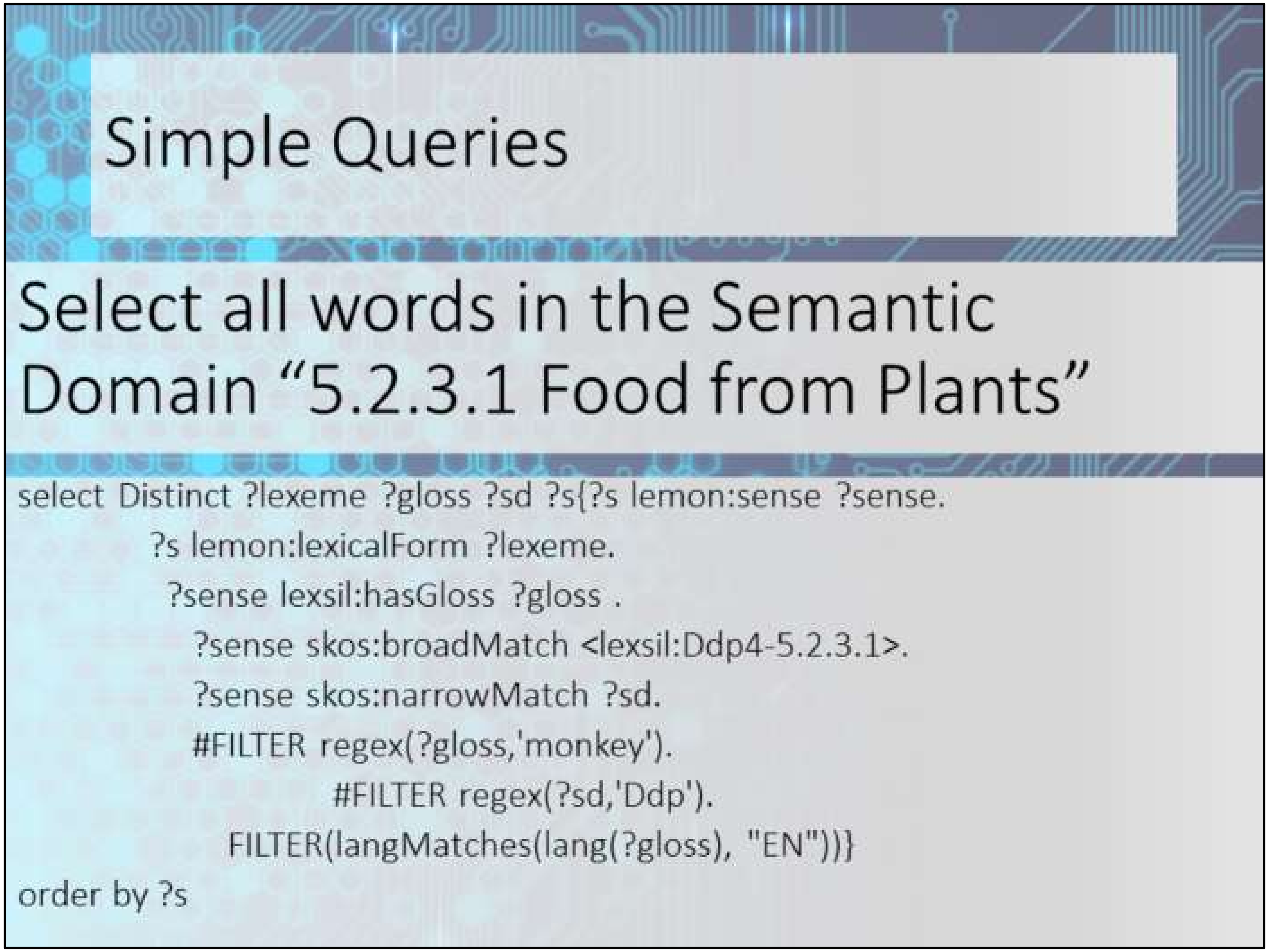
What are all the words that have a gloss or definition including “moon”?

```
select Distinct ?lx ?def ?gloss ?s where {
{ ?ent lemon:lexicalForm  ?lx .
  ?ent lemon:sense  ?s.
  ?s lexsil:hasGloss  ?gloss
FILTER regex(?gloss,'^(lune|moon)$')  }
UNION
{ ?ent lemon:lexicalForm  ?lx .
 ?ent lemon:sense  ?s.
 ?s lexsil:hasDefinition  ?def
FILTER regex(?def,'(lune|moon)(  |$)')    } }
Order by ?s
```



Simple Query (moon/lune)

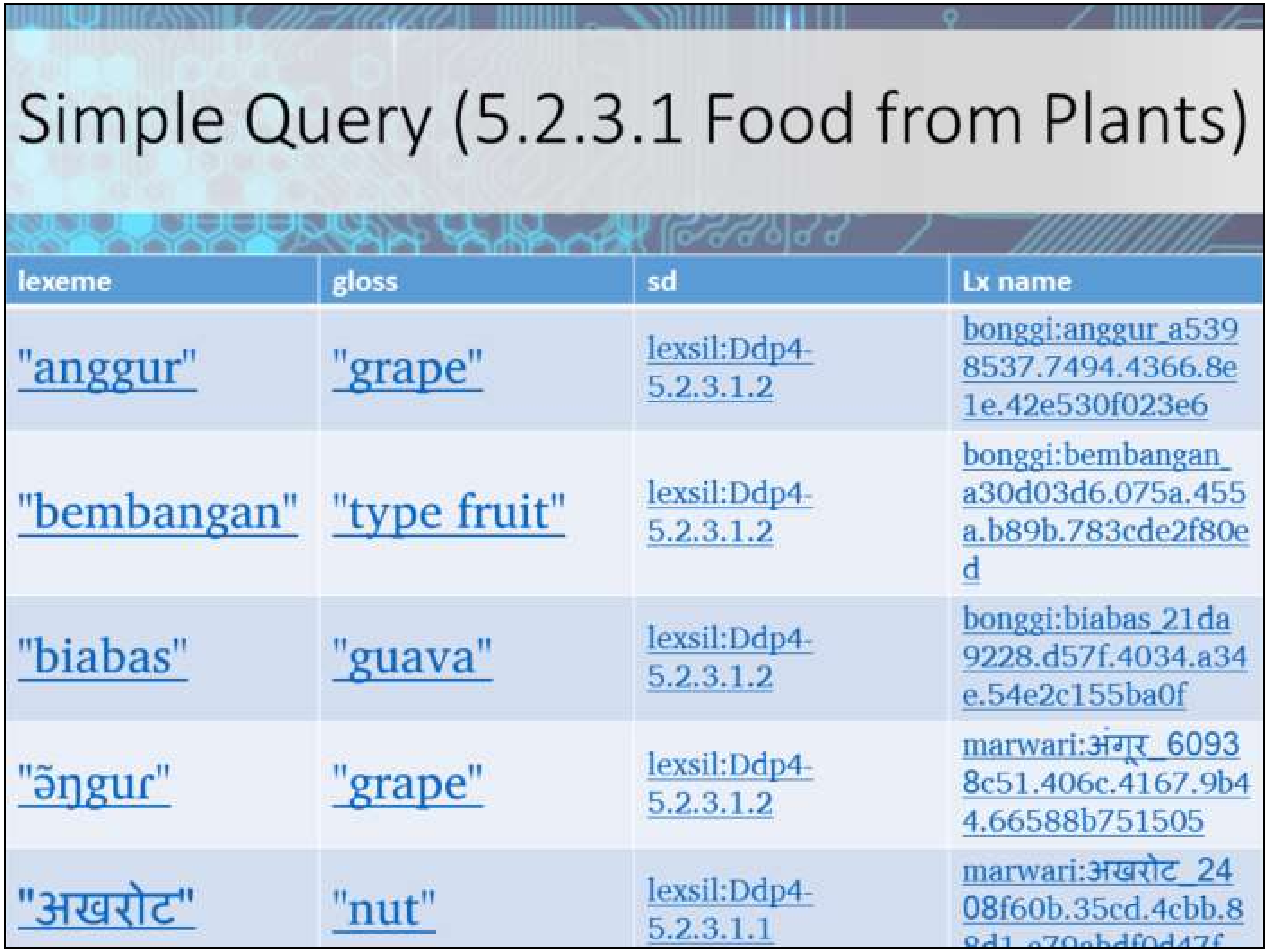
lx	def	gloss	Lang
"lomo"		"lune"	Badwee
"buaidn"		"moon"	Bonggi
"mata-buaidn"		"moon"	Bonggi
"lugut"	"clouds covering moon or stars"		Bonggi
"kelomon"	"very dark; no moon"		Bonggi
"ǹkos"	"Bâton-talisman, bâtonnet magique, baguette de fée utilisée pour rendre un culte à la nouvelle lune. "		Ewondo
"məmua"	"Pleine lune"		Ewondo



Simple Queries

Select all words in the Semantic Domain “5.2.3.1 Food from Plants”

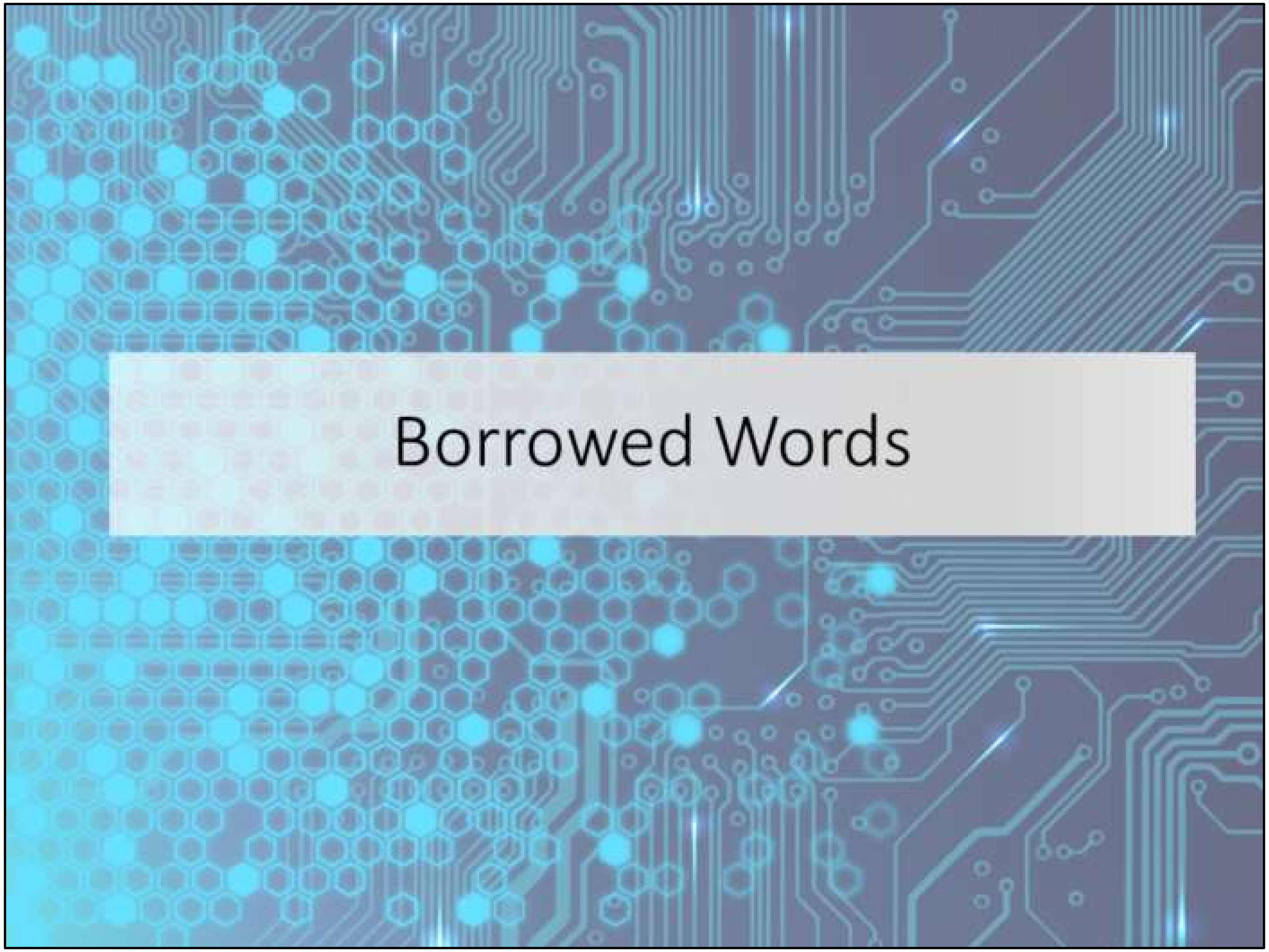
```
select Distinct ?lexeme ?gloss ?sd ?s{?s lemon:sense ?sense.
          ?s lemon:lexicalForm ?lexeme.
           ?sense lexsil:hasGloss ?gloss .
             ?sense skos:broadMatch <lexsil:Ddp4-5.2.3.1>.
             ?sense skos:narrowMatch ?sd.
             #FILTER regex(?gloss,'monkey').
                        #FILTER regex(?sd,'Ddp').
               FILTER(langMatches(lang(?gloss), "EN"))}
order by ?s
```



Simple Query (5.2.3.1 Food from Plants)

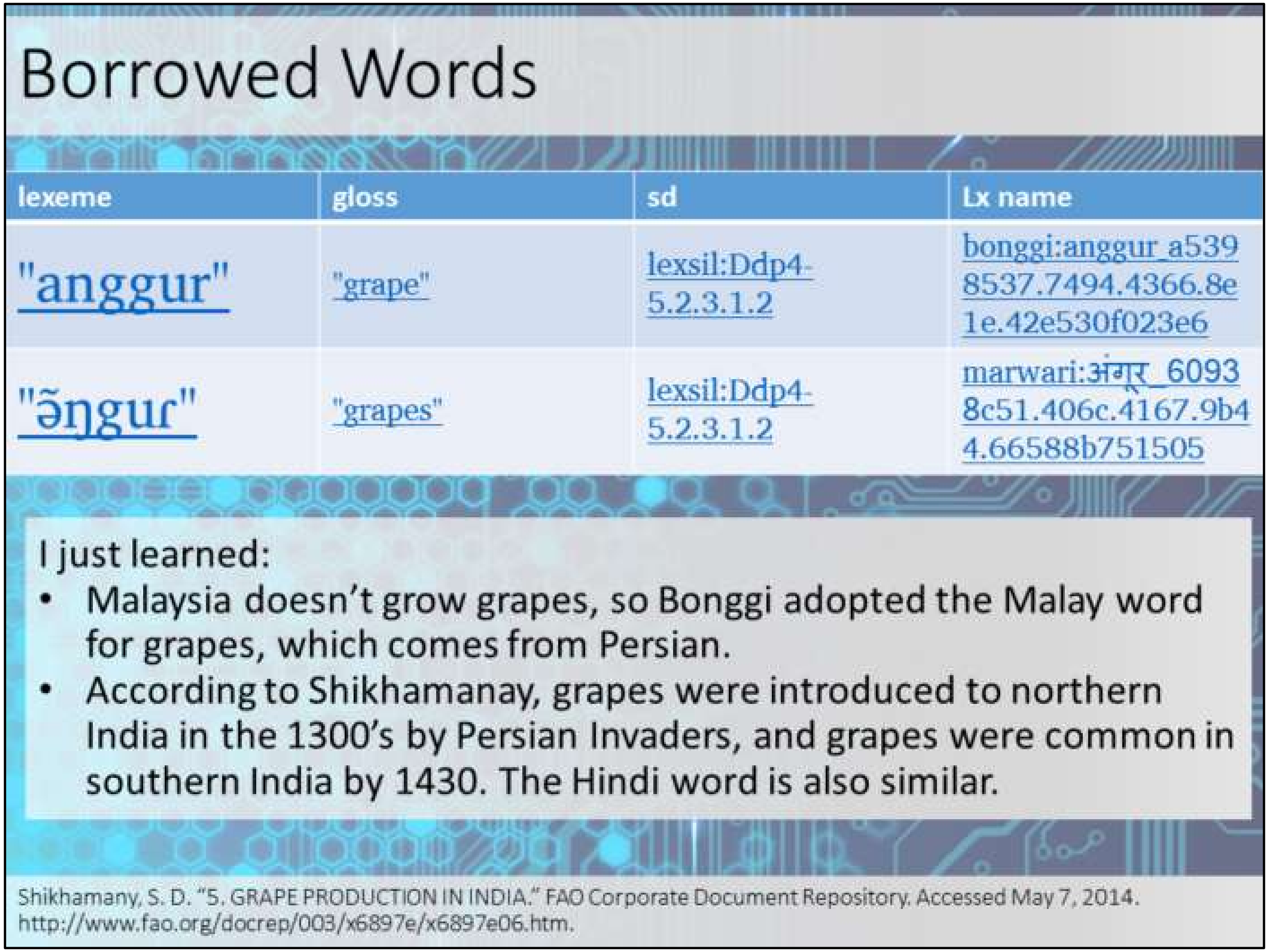
lexeme	gloss	sd	Lx name
"anggur"	"grape"	lexsil:Ddp4-5.2.3.1.2	bonggi:anggur_a5398537.7494.4366.8e1e.42e530f023e6
"bembangan"	"type fruit"	lexsil:Ddp4-5.2.3.1.2	bonggi:bembangan_a30d03d6.075a.455a.b89b.783cde2f80ed
"biabas"	"guava"	lexsil:Ddp4-5.2.3.1.2	bonggi:biabas_21da9228.d57f.4034.a34e.54e2c155ba0f
"ə̃ŋgur"	"grape"	lexsil:Ddp4-5.2.3.1.2	marwari:अंगूर_60938c51.406c.4167.9b44.66588b751505
"अखरोट"	"nut"	lexsil:Ddp4-5.2.3.1.1	marwari:अखरोट_2408f60b.35cd.4cbb.88d1.e79abdf0d47f

Notice that I have asked for "broad" matches. Semantic domains are hierarchical, so anything in category 5.2.3.1.2 is also contained one step up in 5.3.2.1.



Borrowed Words

I didn’t have any related lexicons, and the lexicons I had didn’t have much information on borrowing...so I thought that I wouldn’t be able to show this with my dataset. But…



Borrowed Words

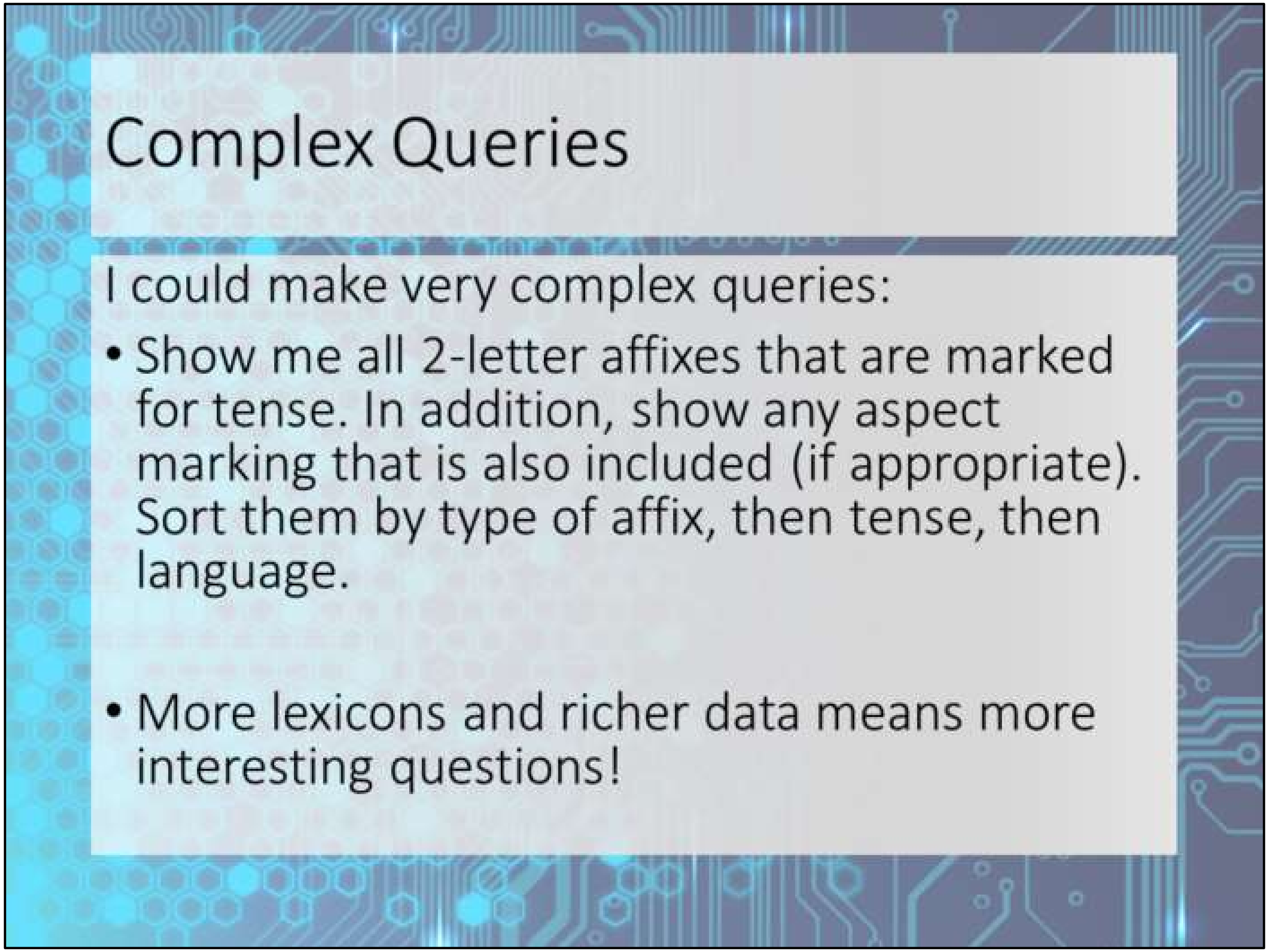
lexeme	gloss	sd	Lx name
"anggur"	"grape"	lexsil:Ddp4-5.2.3.1.2	bonggi:anggur_a5398537.7494.4366.8e1e.42e530f023e6
"ə̃ŋgur"	"grapes"	lexsil:Ddp4-5.2.3.1.2	marwari:अंगूर_60938c51.406c.4167.9b44.66588b751505

I just learned:

- Malaysia doesn't grow grapes, so Bonggi adopted the Malay word for grapes, which comes from Persian.
- According to Shikhamanay, grapes were introduced to northern India in the 1300's by Persian Invaders, and grapes were common in southern India by 1430. The Hindi word is also similar.

Shikhamany, S. D. "5. GRAPE PRODUCTION IN INDIA." FAO Corporate Document Repository. Accessed May 7, 2014. http://www.fao.org/docrep/003/x6897e/x6897e06.htm.

Take a look at the words for grapes from the previous search… This is part of the fun of cross-linguistic study.

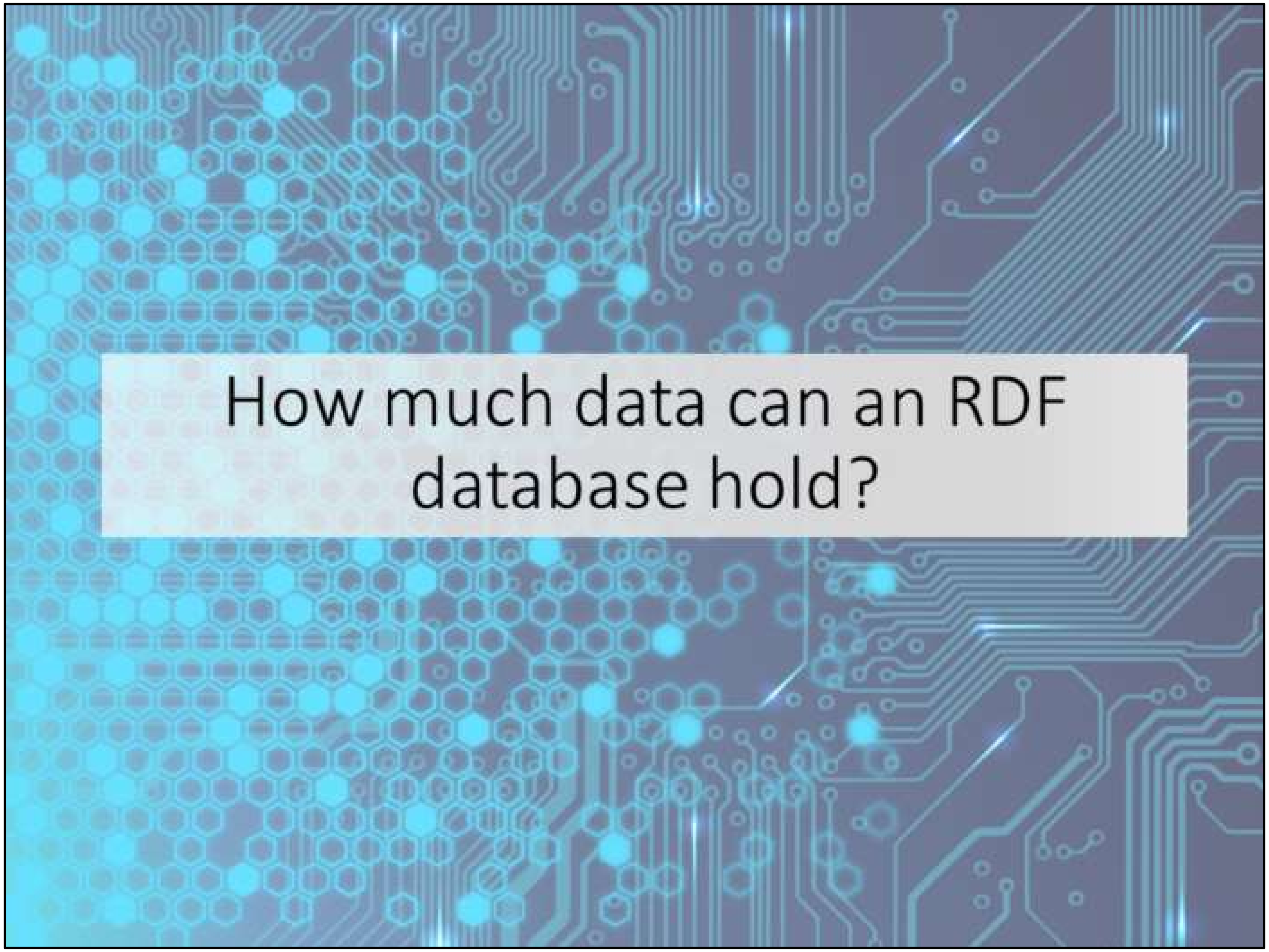


Complex Queries

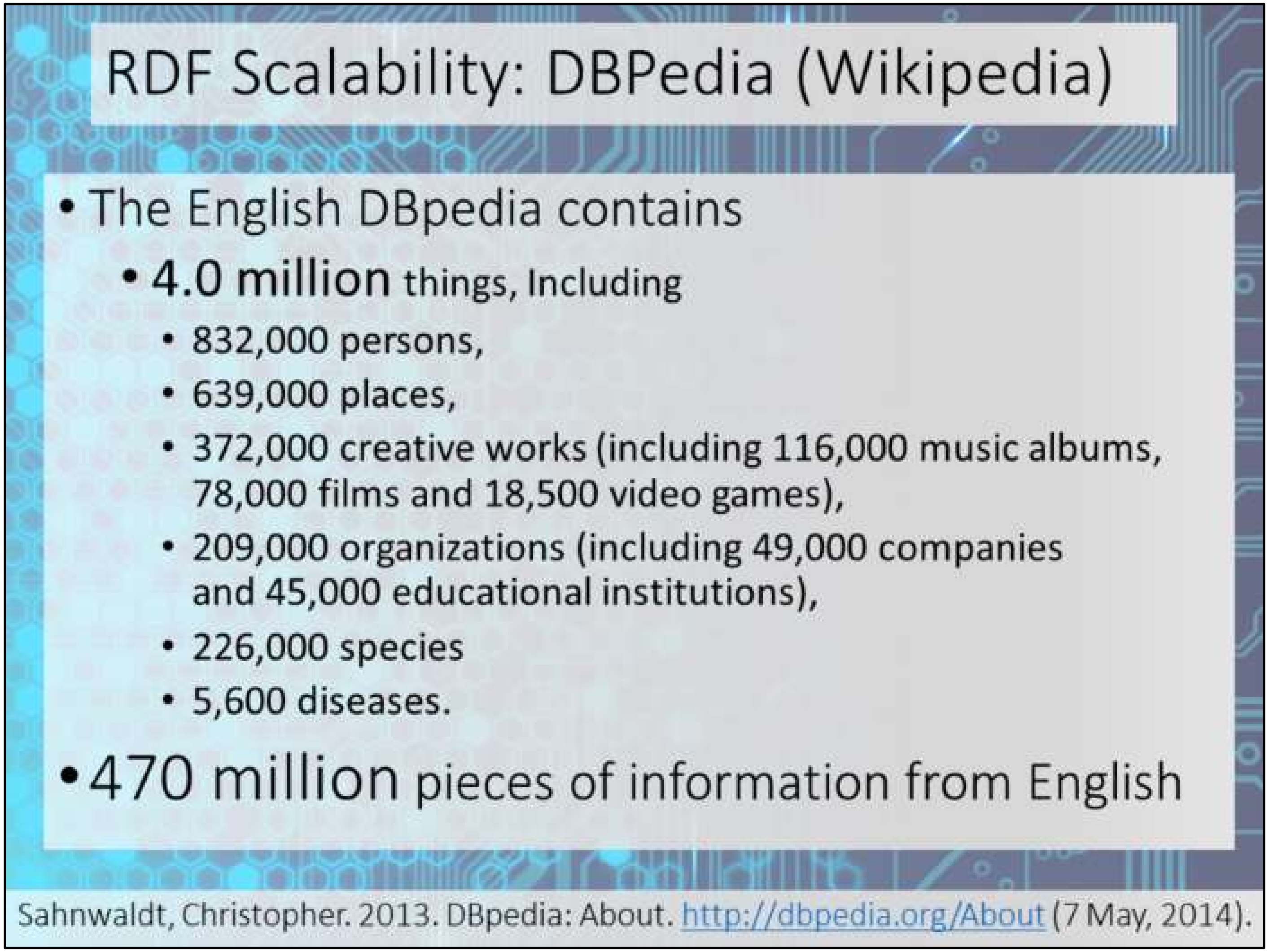
I could make very complex queries:

- Show me all 2-letter affixes that are marked for tense. In addition, show any aspect marking that is also included (if appropriate). Sort them by type of affix, then tense, then language.

- More lexicons and richer data means more interesting questions!



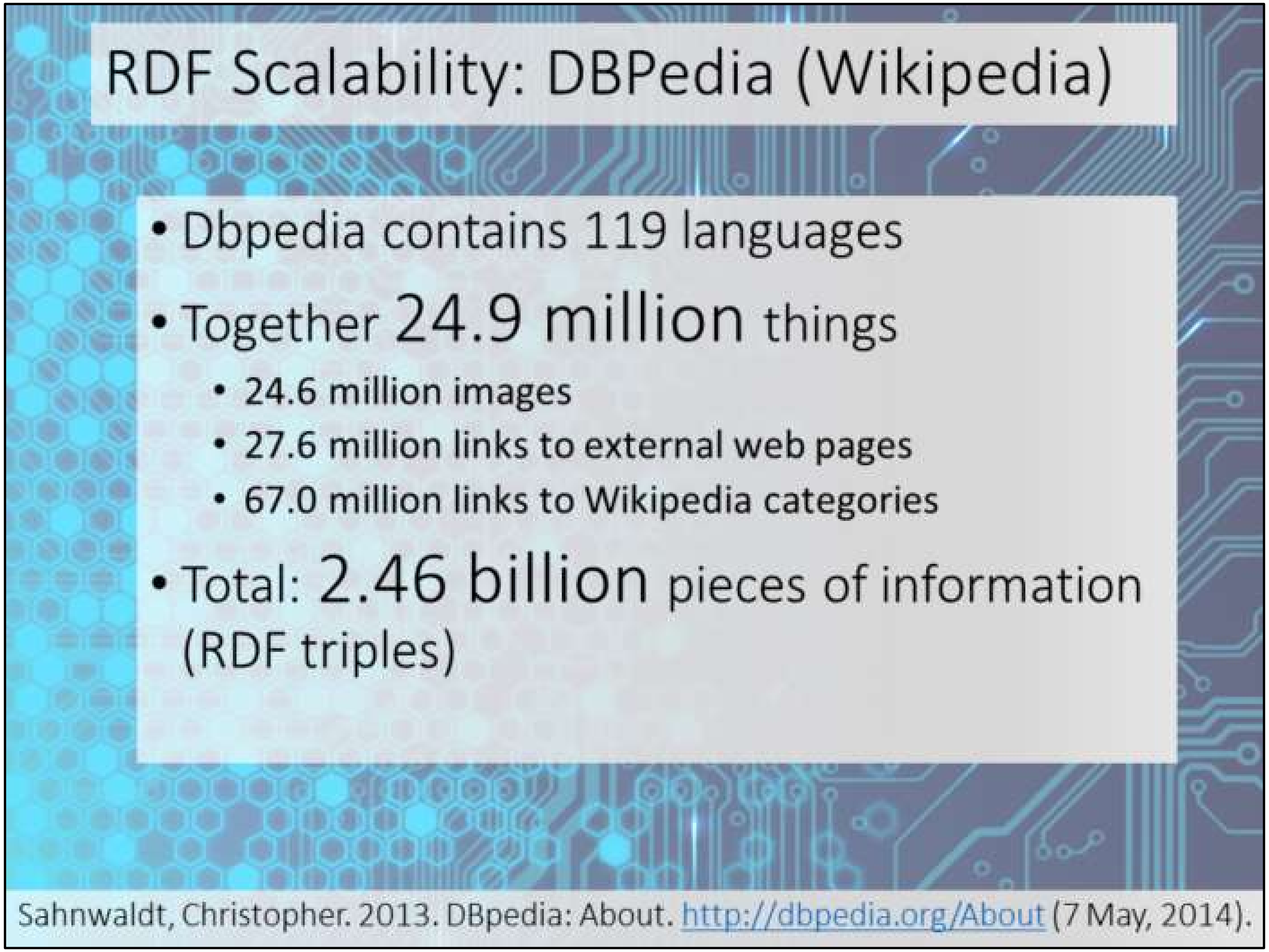
How much data can an RDF database hold?



RDF Scalability: DBPedia (Wikipedia)

- The English DBpedia contains
  - 4.0 million things, Including
    - 832,000 persons,
    - 639,000 places,
    - 372,000 creative works (including 116,000 music albums, 78,000 films and 18,500 video games),
    - 209,000 organizations (including 49,000 companies and 45,000 educational institutions),
    - 226,000 species
    - 5,600 diseases.
- 470 million pieces of information from English

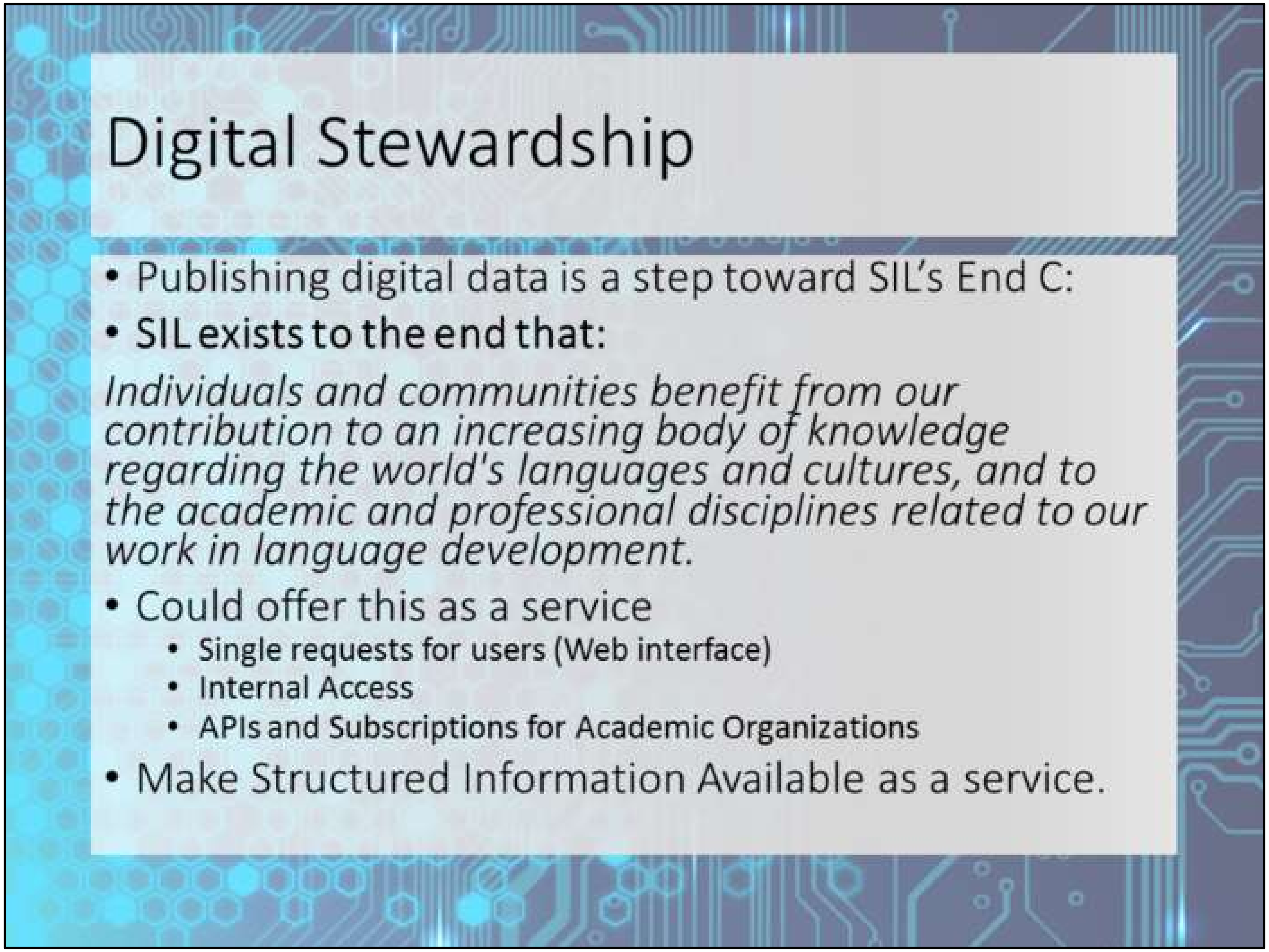
Sahnwaldt, Christopher. 2013. DBpedia: About. http://dbpedia.org/About (7 May, 2014).



RDF Scalability: DBPedia (Wikipedia)

- Dbpedia contains 119 languages
- Together 24.9 million things
    - 24.6 million images
    - 27.6 million links to external web pages
    - 67.0 million links to Wikipedia categories
- Total: 2.46 billion pieces of information (RDF triples)

Sahnwaldt, Christopher. 2013. DBpedia: About. http://dbpedia.org/About (7 May, 2014).

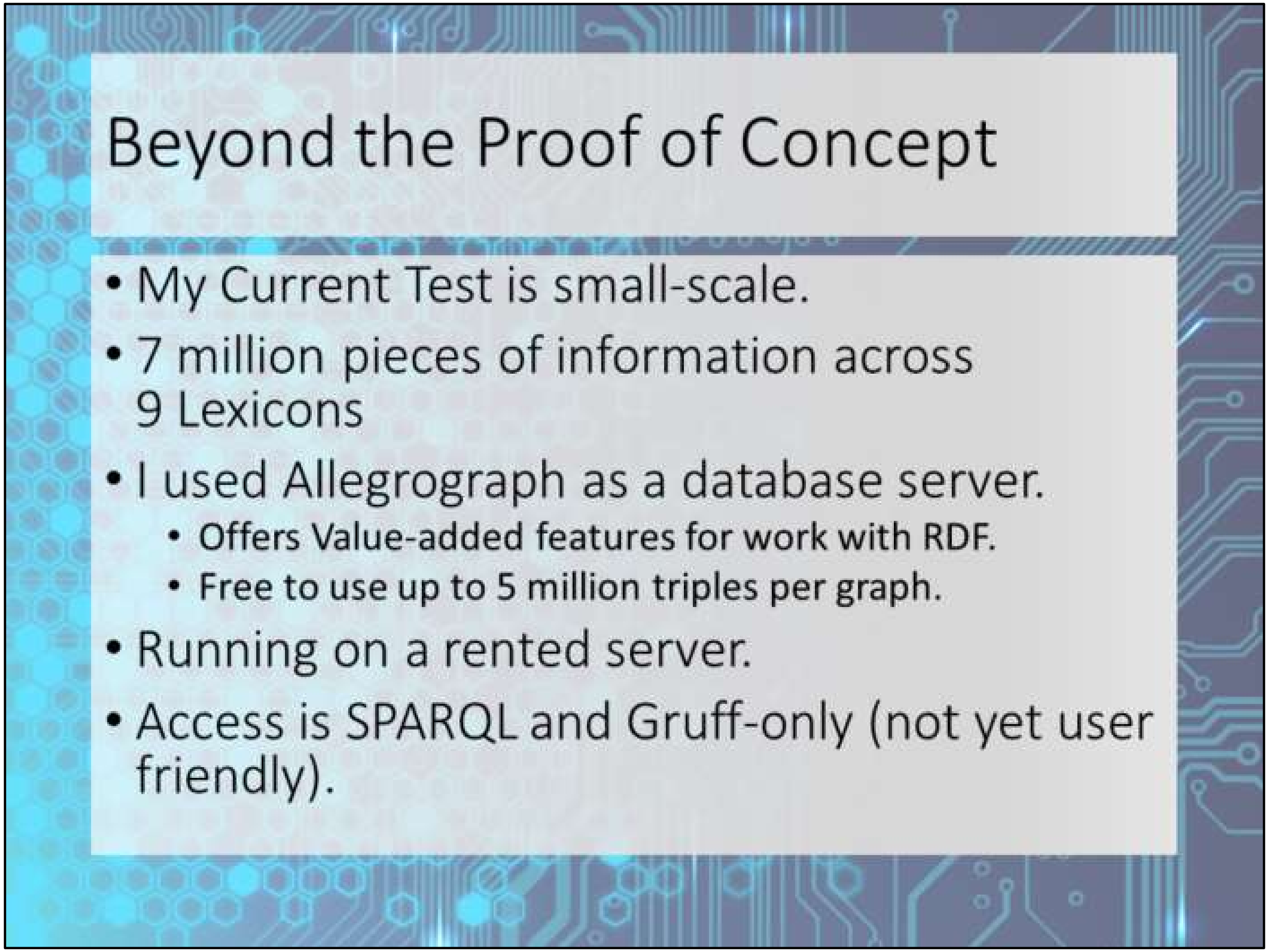


Digital Stewardship

- Publishing digital data is a step toward SIL's End C:
- SIL exists to the end that:

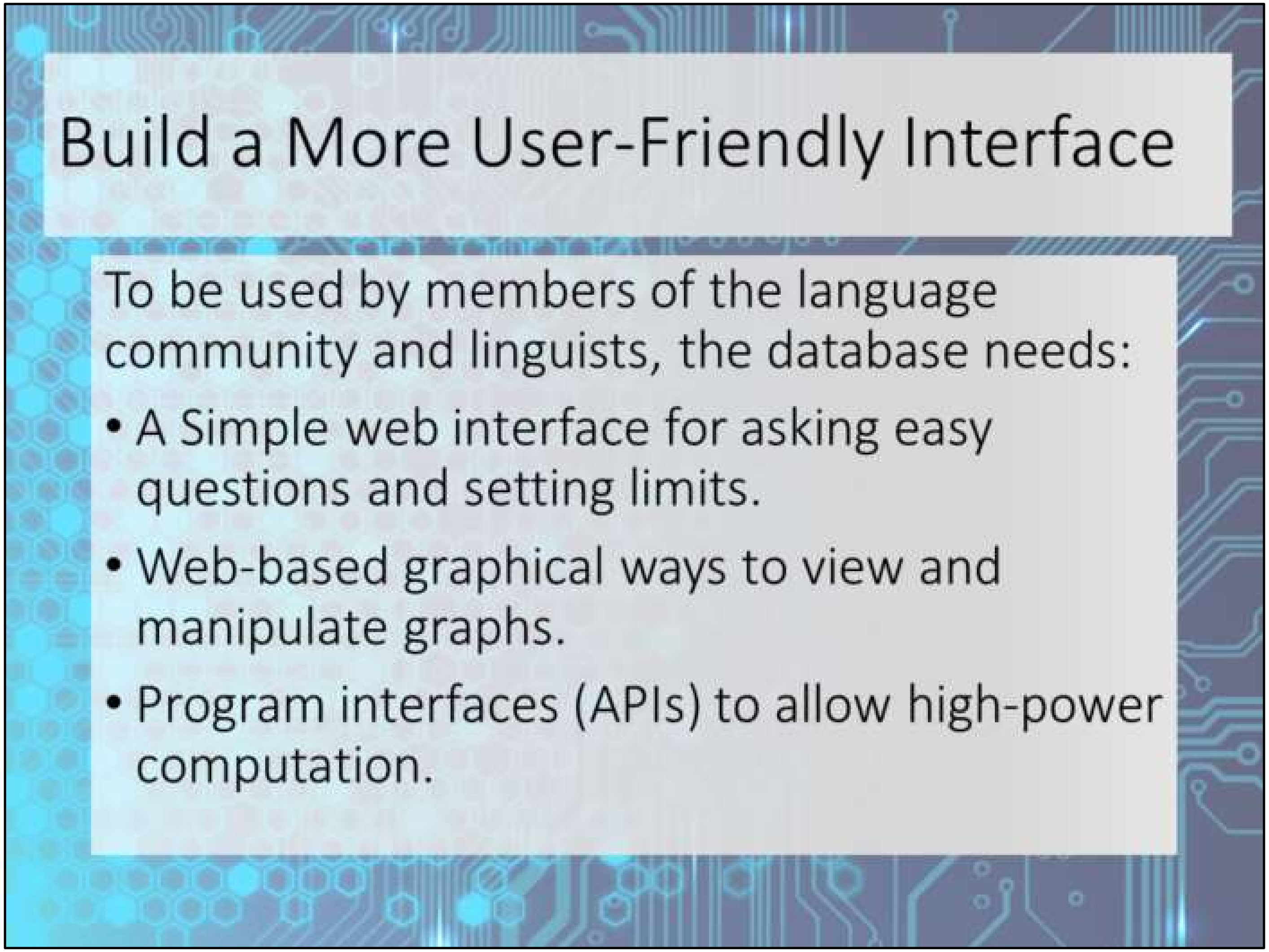
Individuals and communities benefit from our contribution to an increasing body of knowledge regarding the world's languages and cultures, and to the academic and professional disciplines related to our work in language development.

- Could offer this as a service
  - Single requests for users (Web interface)
  - Internal Access
  - APIs and Subscriptions for Academic Organizations
- Make Structured Information Available as a service.



Beyond the Proof of Concept

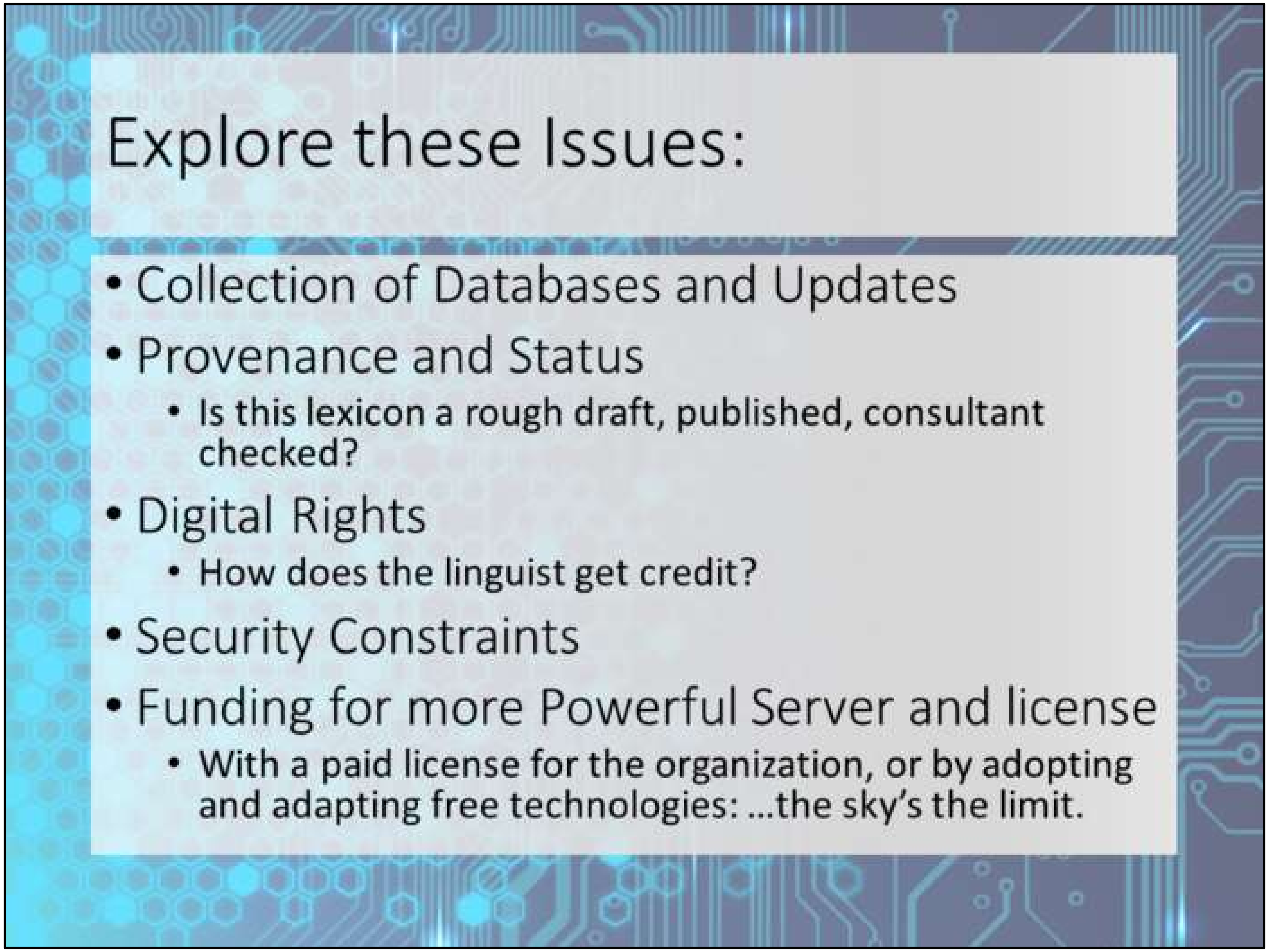
- My Current Test is small-scale.
- 7 million pieces of information across 9 Lexicons
- I used Allegrograph as a database server.
  - Offers Value-added features for work with RDF.
  - Free to use up to 5 million triples per graph.
- Running on a rented server.
- Access is SPARQL and Gruff-only (not yet user friendly).



Build a More User-Friendly Interface

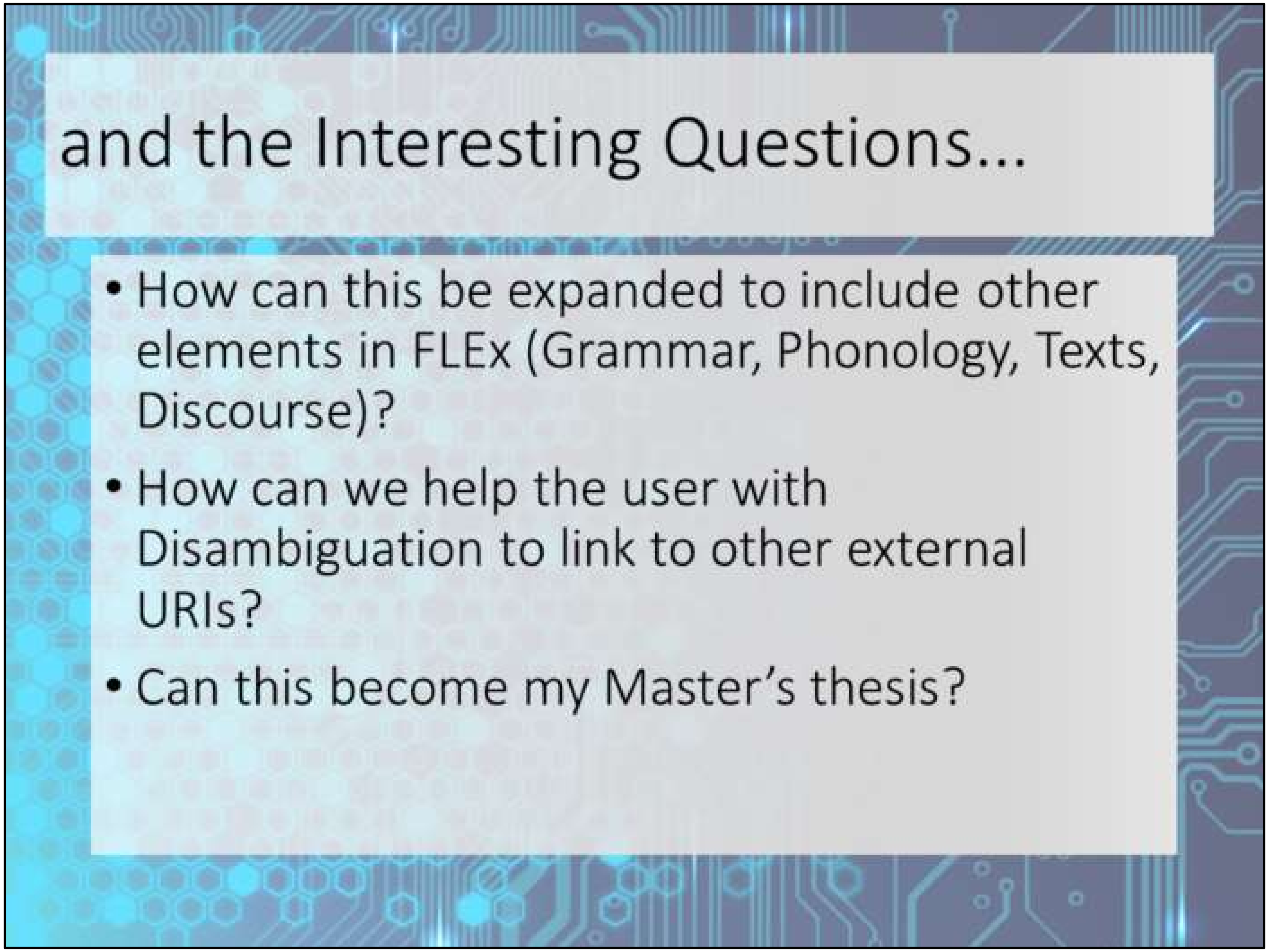
To be used by members of the language community and linguists, the database needs:

- A Simple web interface for asking easy questions and setting limits.
- Web-based graphical ways to view and manipulate graphs.
- Program interfaces (APIs) to allow high-power computation.



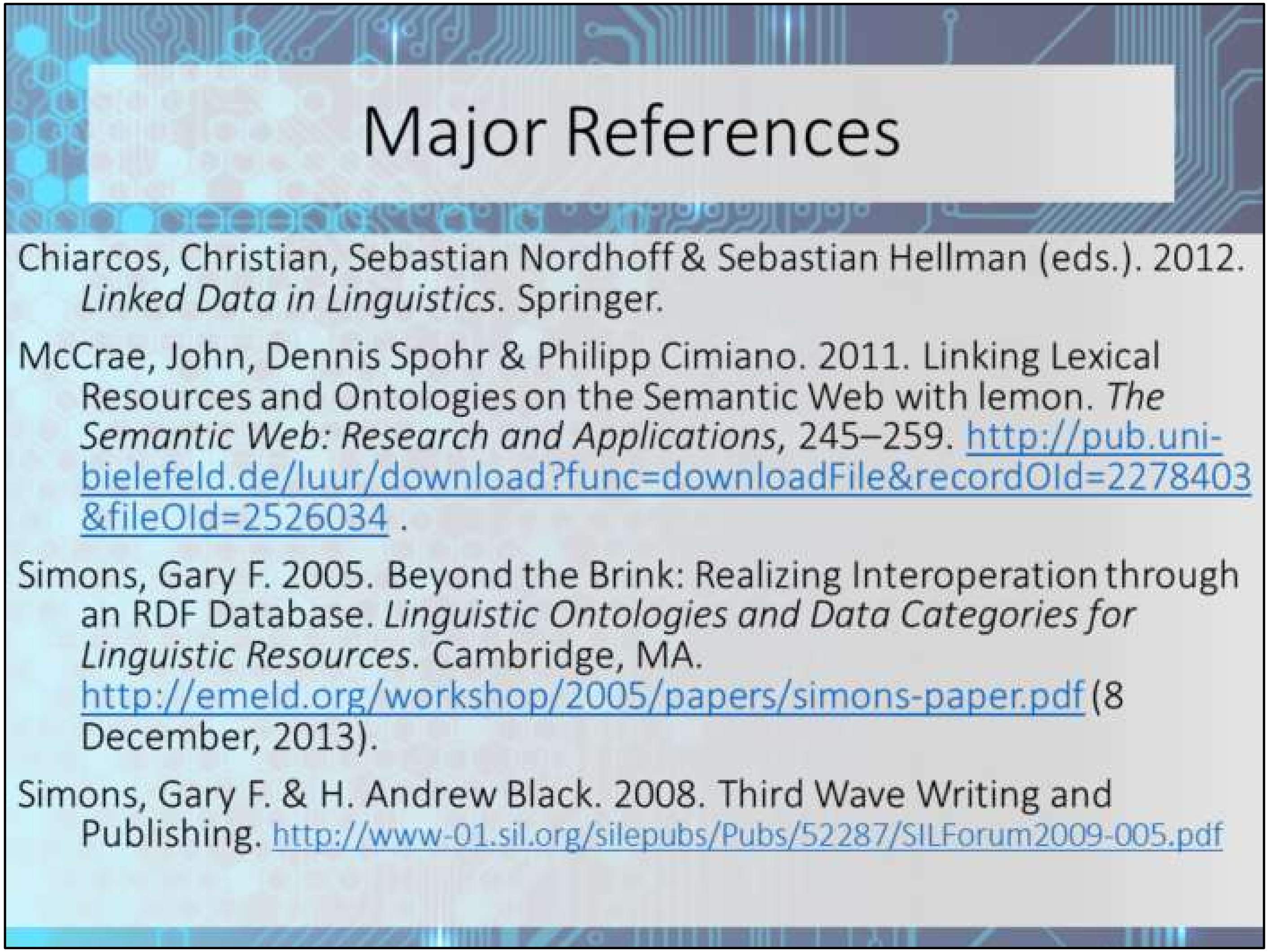
Explore these Issues:

- Collection of Databases and Updates
- Provenance and Status
  - Is this lexicon a rough draft, published, consultant checked?
- Digital Rights
  - How does the linguist get credit?
- Security Constraints
- Funding for more Powerful Server and license
  - With a paid license for the organization, or by adopting and adapting free technologies: …the sky's the limit.



and the Interesting Questions...

- How can this be expanded to include other elements in FLEx (Grammar, Phonology, Texts, Discourse)?
- How can we help the user with Disambiguation to link to other external URIs?
- Can this become my Master's thesis?



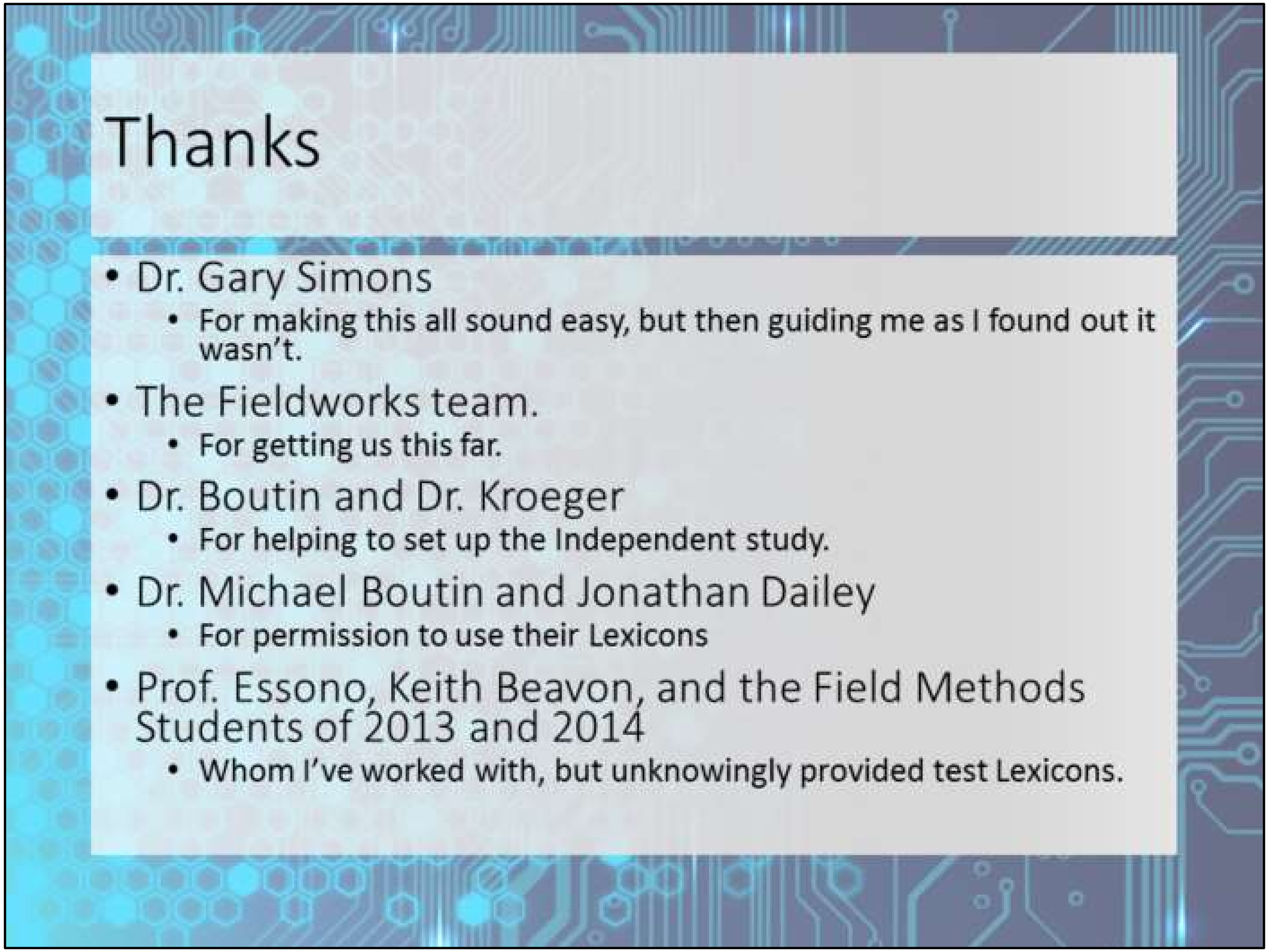
Major References

Chiarcos, Christian, Sebastian Nordhoff & Sebastian Hellman (eds.). 2012. *Linked Data in Linguistics*. Springer.

McCrae, John, Dennis Spohr & Philipp Cimiano. 2011. Linking Lexical Resources and Ontologies on the Semantic Web with lemon. *The Semantic Web: Research and Applications*, 245–259. http://pub.uni-bielefeld.de/luur/download?func=downloadFile&recordOId=2278403&fileOId=2526034 .

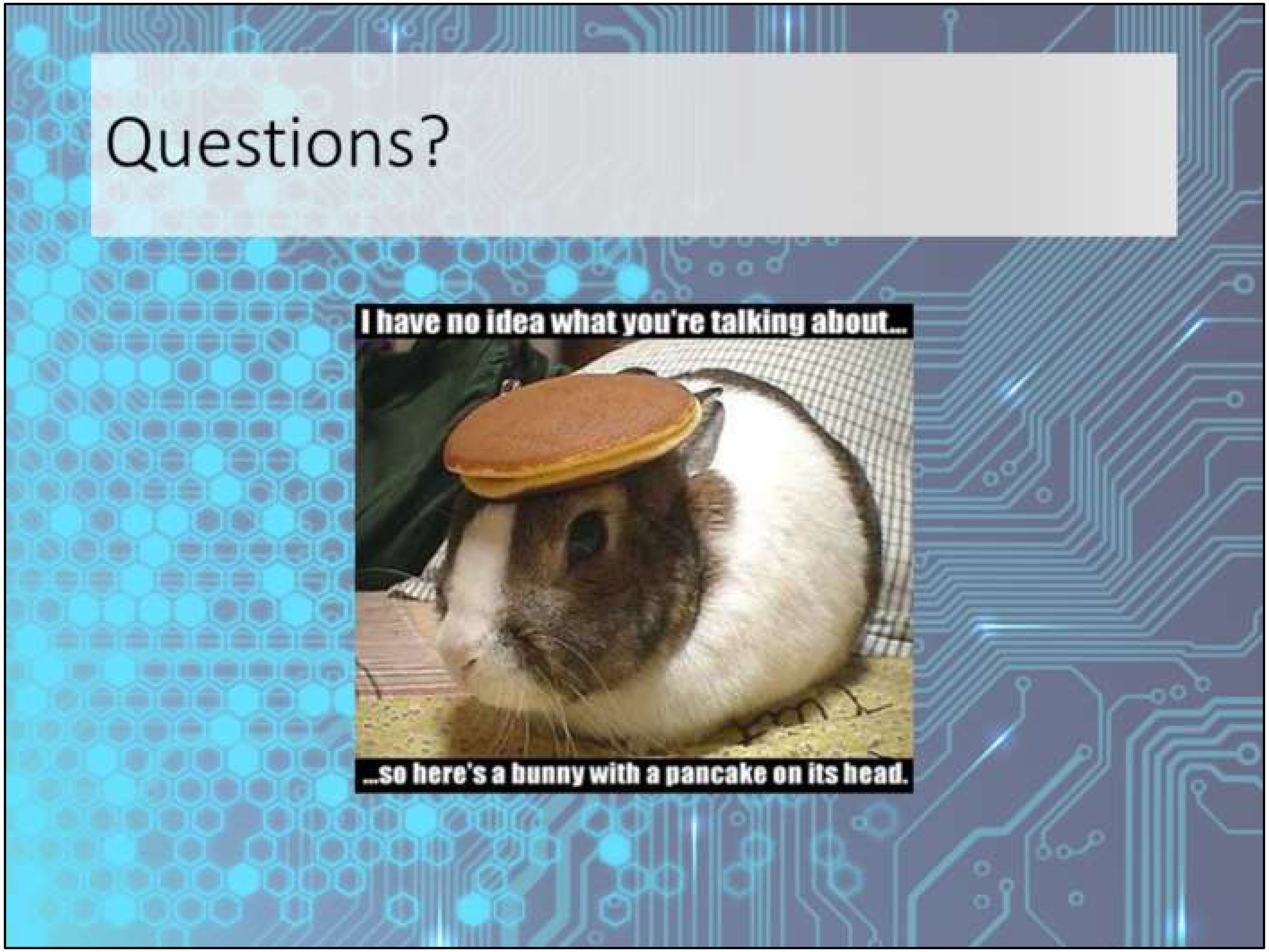
Simons, Gary F. 2005. Beyond the Brink: Realizing Interoperation through an RDF Database. *Linguistic Ontologies and Data Categories for Linguistic Resources*. Cambridge, MA. http://emeld.org/workshop/2005/papers/simons-paper.pdf (8 December, 2013).

Simons, Gary F. & H. Andrew Black. 2008. Third Wave Writing and Publishing. http://www-01.sil.org/silepubs/Pubs/52287/SILForum2009-005.pdf



Thanks

- Dr. Gary Simons
  - For making this all sound easy, but then guiding me as I found out it wasn’t.
- The Fieldworks team.
  - For getting us this far.
- Dr. Boutin and Dr. Kroeger
  - For helping to set up the Independent study.
- Dr. Michael Boutin and Jonathan Dailey
  - For permission to use their Lexicons
- Prof. Essono, Keith Beavon, and the Field Methods Students of 2013 and 2014
  - Whom I’ve worked with, but unknowingly provided test Lexicons.



Questions?

I have no idea what you're talking about...
...so here's a bunny with a pancake on its head.