# Termbase eXchange (TBX) Making Exchange Work for You

**Dr. Sue Ellen Wright**

Kent State University

March 18, 2018

Association for Machine Translation in the Americas

Boston, MA

# Fragmentation, Heterogeneity, and Non-Interoperability

- Fragmentation, heterogeneity and a lack of interoperability between methods, tools and data sets (Jen's first slide)
- Applicable for termbase design as well
- Issues
  - TBX only viable for coherent data models
  - Prevalence of non-complying models
  - Lack of guidance regarding viable models
  - Need for coordination with XLIFF
  - Need for xmlns documentation
  - Outdated link handling

# Increasing Collaboration of MT and CAT

- Agile combinations and recombinations of MT and Computer Assisted Translation (CAT)—difficult to separate out approaches except for limited evaluation of tools
  - MT as an option in CAT
  - Predictive MT in CAT
  - Role of terminology management in governing human interaction with MT and TM
  - Potential issue: interface between TBX & LMF

# ISO 30042 – TermBase eXchange

- Issued in 2009, but subject to ongoing development
- New Draft International standard currently in ballot
- Archiving the information in a termbase
- Exchanging information between systems
  - Authoring (send monolingual information from a termbase to an authoring tool)
  - Translation (send a subset of the information from a termbase to a translator)
  - Data mining (export most/all information from a termbase for analysis using XML)
- Guiding the design of a new termbase for interoperability

# ISO 30042 :2009– TermBase eXchange

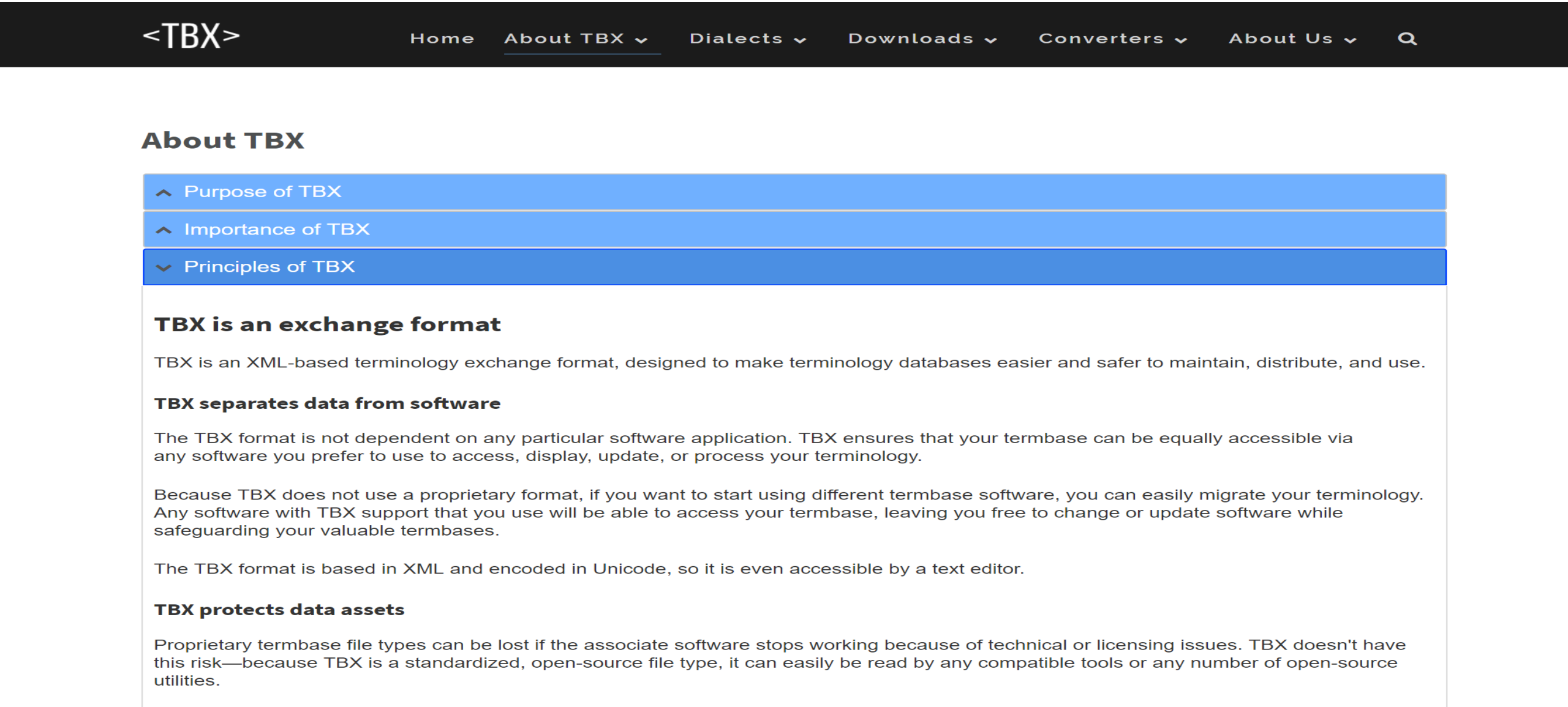
- Positives
  - Powerful and flexible enough to express almost any termbase model
  - Viable for exchange among like data models & systems
- Negatives
  - Not widely used for exchange between divergent systems
  - Applications outputs not true TBX
  - TBX as a moving target
  - Exchange impossible between incompatible data models
  - Incompatibility with some modern xml solutions

# TBX-Basic

- The localization industry solution to a normalized format
- "Dialect" of TBX consisting of a defined data model and a specified set of data categories
- Positive:
  - Standardized data model & selection of data categories (datcats)
  - Recognized user community
- Negative:
  - Rich, but restricted data category set
  - Need for reliable support tools
  - Some outdated XML conventions

| TBX-Basic Dialect |
|---|
| ***Concept Level*** |
| Subject field |
| Image |
| Note |
| Definition |
| Source of definition |
| Cross-reference/Reference |
| Creation Date |
| Created by |
| Last Modified Date |
| Last Modified by |
| |
| ***Language Section*** |
| Language |
| Definition |
| Source |
| Note |
| Creation Date |
| Created by |
| Last Modified Date |
| Last Modified by |
| |
| ***Term Section*** |
| Term |
| Source |
| Part of Speech |
| Gender |
| Usage Status |
| Term Type |
| Geographical Usage |
| Context |
| Source of context |
| Note |
| Cross-reference/Reference |
| Term Location |
| Customer |
| Project |
| External Cross-Reference |
| |

# Support tools: tbxInfo.net




TBXinfo.net screenshots with permission of LTAC Global

<TBX> Home DatCats Modules Dialects Downloads Tools About

Master List

View All
TBX-Core
TBX-Min
TBX-Basic
Creating Dialects
Validating Dialects
Private Dialects

View All
MRC to TBX-Basic
MultiTerm XML – TBX-Default
Spreadsheet Glossary – TBX-Min
TBX-Basic – TBX-Min
UTX – TBX-Min

# ISO 30042 Ballot

The production version of the TBX website (http://www.tbxinfo.net) wil
closes on April 26th. During the time, please see the development vers
be implemented. By March 1st, everyone interested in TBX will be able
website comments will be separate from the official ISO commenting t

Any mention of **TBX Default** should be understood to refer to the master list of TBX data categories, rather than to a TBX dialect. Eventually (after the ballot ends in April), the website will be updated and the obsolete term TBX Default will no longer be mentioned except in historical notes.

---

**TBX**, or **TermBase eXchange**, is the international standard for *representing* and *exchanging* information about terminology.

The current version of the TBX standard was published in 2008. The next version is under development. In preparation for the next version, whenever you receive a TBX file, please check the value of the type attribute on the root element.

For example, in <martif type="TBX" xml:lang="en"> the type is simply TBX. The constraints on TBX are expected to be in an XCS file. However, in practice, the XCS file is often missing or not processed.

In the next version of TBX, the root element will be "tbx" instead of "mar
dialect, for example "TBX-Basic". Each dialect name is associated with a
modules. Each module clearly indicate which data categories are allow

http://www.tbxinfo.net

This change will address the single most common complaint about the current version of TBX: if there is no XCS file associated with a TBX document instance, you don't know what to expect. In the new version there is no XCS file; there is a dialect name; and you do know what to expect.

# Upgrading and Empowering TBX

- ISO 30042:2018 and beyond
- Coordination with XLIFF terminology markup
- Use of xml namespaces
- Modernizing hypertext representations
  - TEI Term (ancestor of TBX) predated HTML & modern idref/href notations
  - /cross-reference/ envisioned as a datcat
  - Enabled by its own linking features

```
<descripSpec name="definition" datcatId="ISO12620A-0501">
    <contents/>
    <levels>langSet termEntry</levels>
</descripSpec>
<xrefSpec name="externalCrossReference" datcatId="ISO12620A-101807">
    <contents targetType="external"/>
</xrefSpec>
```

# When TBX-Basic is not enough

- TBX-Linguist
  - Additional data categories
  - /figure/ added at the language level
  - /cross-reference/ swapped out for /related concept/ and /related term/
  - /register/
  - /grammaticalNumber/
  - /transferComment/
  - ja-specific datcats (reading, readingNote)

<TBX>

# TBX-Linguist

| TBX-Basic | TBX-Linguist |
|---|---|
| *Concept Level* | |
| Subject Field | Subject Field |
| | Entry Identifier |
| Figure * | Figure* |
| Source of Figure* | Source |
| Note | Note |
| | Source |
| Definition | Definition |
| Source of Definition | Source |
| Related Concept | Related Concept |
| Customer Subset | Customer Subset |
| Project Subset | Project Subset |
| Created by | Created by |
| Last Modified Date | Last Modified Date |
| Last Modified by | Last Modified by* |
| External Cross-Reference | External Cross-Reference |
| *Language Section* | |
| Language | Language |
| Figure* | Figure* |
| Source of Figure* | Source |
| Definition | Definition |
| Source | Source |
| Note | Note |
| Source of Note* | Source |
| [Transaction Set] | [Transaction Set] |
| | |

Filter/sorting fields

| TBX-Basic | TBX-Linguist |
|---|---|
| *Term Section* | |
| Term | Term |
| Source | Source |
| Part of Speech | Part of Speech |
| Gender | Gender |
| | Reading |
| | Reading Note |
| | Grammatical Number |
| Administrative Status | Term Status |
| Term Type | Term Type |
| Geographical Usage | Geographical Usage |
| | Usage Register |
| Context | Context |
| Source of Context | Source of context |
| Note | Note |
| Source of Note* | N-Source |
| Related Term | Related Term |
| | Transfer Comment |
| Term Location | Term Location |
| Customer Subset | Customer Subset |
| Project Subset | Project Subset |
| External Cross-Reference | External Cross-Reference |
| [Transaction Set] | [Transaction Set] |

Language specific fields

# When TBX-Basic is not enough

- TBX-Linguist
  - Additional data categories
  - /figure/ added at the language level
  - /cross-reference/ swapped out for /related concept/ and /related term/
  - /register/
  - /grammaticalNumber/
  - /transferComment/
  - ja-specific datcats (reading, readingNote)

Entry number: 22

German

figure:


source: http://www.katzen-album.de/forum/viewtopic.php?t=24127 2017-07-13

definition: eine Familie (Félidae) der Raubtiere

source: http://www.wortbedeutung.info/Katze/ 2017-07-13

Term: Katze

partOfSpeech: noun

administrativeStatus: preferred

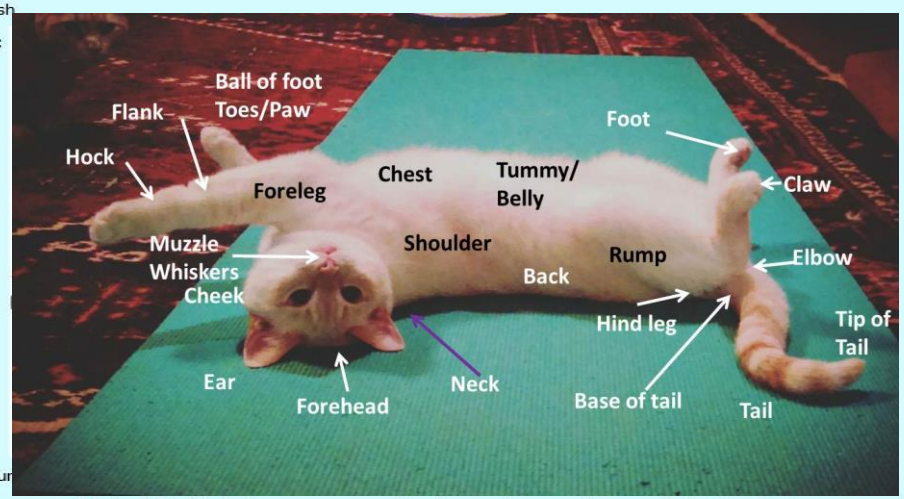
grammaticalGender: feminine

context: Mit ihren scharfen Krallen können Katzen sehr gut Bäume hochklettern, aber zum Abstieg muss die Katze gelernt haben, ihre nach vorne gekrümmten Krallen als „Steighaken" zu benutzen. Unerfahrene Katzen versuchen, mit dem Kopf voraus nach unten zu klettern, wobei sie schnell in Schwierigkeiten kommen können, in Panik geraten und in eine Schockstarre verfallen.

source: https://de.wikipedia.org/wiki/Katzen 2017-07-13

English

figure:




sour

definition: any of a family (Felidae) of carnivorous usually solitary and nocturnal mammals (such as the domestic cat, lion, tiger, leopard, jaguar, cougar, wildcat, lynx, and cheetah)

# Figures at the Language Level

# Term Entry Links

Entry level

Entry number: 1

subjectField: Biologie

figure:

source: http://www.mirror.co.uk/news/weird-news/devil-cat-shiny-terrorised-hospitalised-2893846, 2017-07-11

crossReference: Entry 5 Hauskatze

German

definition: männliche Hauskatze

source: https://de.wikipedia.org/wiki/Hauskatze, 2017-07-11

Term: Kater

partOfSpeech: noun

context: Kater haben zusätzlich eine Anhäufung von Duftdrüsen in einer Art mit einem Kanal versehenen Tasche neben dem Anus. Alle Schweiß- und Talgdrüsen dienen hauptsächlich der Kommunikation über den Geruch durch Reiben an Gegenständen, Artgenossen und Personen.

source: https://de.wikipedia.org/wiki/Hauskatze, 2017-07-11

grammaticalGender: masculine

crossReference: Hauskatze

English

definition: adult unneutered male house cat

Term: tomcat

partOfSpeech: noun

termType: fullForm

context: A male cat is called a "tom" or a tomcat.

source: https://en.wikipedia.org/wiki/Cat, 2017-07-11

crossReference: domestic cat

Term: tom

partOfSpeech: noun

termType: shortForm

**1. URIs**

**2. cross-reference to entry**

**3. cross-reference to related term**

*(TBXBasic)*

# Term Entry Links

Entry level

Entry number: 20

subjectField: Biologie

figure:

source: http://www.mirror.co.uk/news/weird-news/devil-cat-shiny-terrorised-hospitalised-2893846, 2017-07-11

relatedEntry: Entry 19 Hauskatze

German

definition: männliche Hauskatze

source: https://de.wikipedia.org/wiki/Hauskatze, 2017-07-11

Term: Kater

partOfSpeech: noun

grammaticalGender: masculine

context: Kater haben zusätzlich eine Anhäufung von Duftdrüsen in einer Art mit einem Kanal versehenen Tasche neben dem Anus. Alle Schweiß- und Talgdrüsen dienen hauptsächlich der Kommunikation über den Geruch durch Reiben an Gegenständen, Artgenossen und Personen.

source: https://de.wikipedia.org/wiki/Hauskatze, 2017-07-11

relatedTerm: Hauskatze

English

definition: adult unneutered male house cat

Term: tomcat

partOfSpeech: noun

termType: fullForm

context: A male cat is called a "tom" or "tomcat.

source: https://en.wikipedia.org/wiki/Cat, 2017-07-11

relatedTerm: domestic cat

Term: tom

1. URIs

2. related entry (concept)

3. related term

*(TBXLinguist)*

# Creating a New Dialect

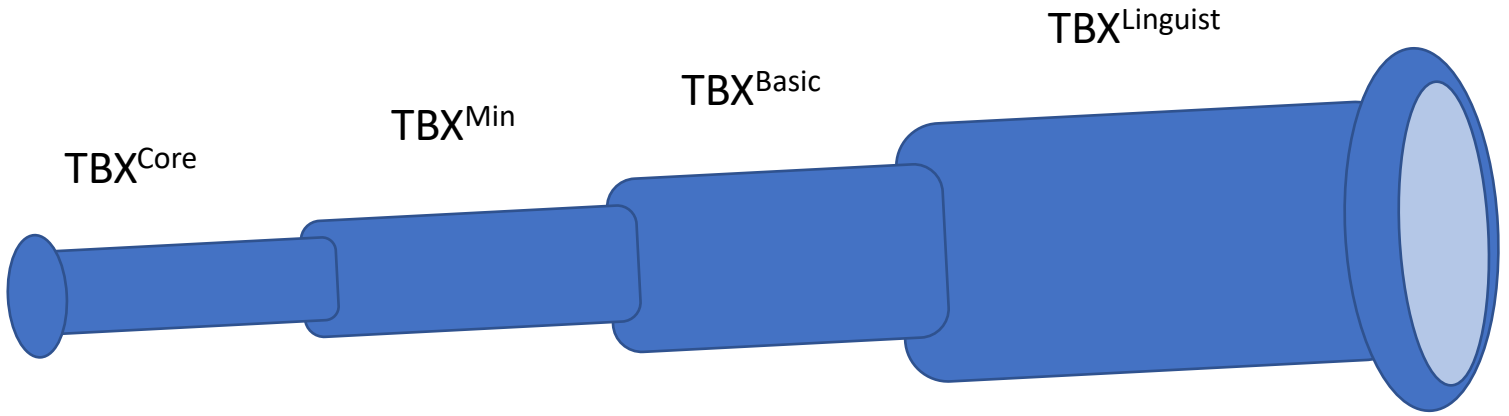
- Start from TBX Basic
  - Even if you don't want to use all of it!
- Add additional DCs to data model
- Map any name changes
- Edit data model for all languages, all terms, and synonyms for each language
- Activate datcats in the respective model levels

# Creating a New Dialect

- Update
  - Data model (previous slide)
  - Layout model (replicate additions at all levels)
  - Input model
    - You can leave out items you don’t want to use in a given iteration
    - But keep them in the core data model & in layout
- Edit TBX-Basic output xml as needed (datcat names)
- Import seamlessly into new data model

# Honor the telescope!

- TBX Core + TBX$^{Min}$ + TBX$^{Basic}$ + TBX$^{Linguist}$ Modules = TBX-Linguist Dialect
- Each successive dialect is a superset of what comes before
- All subordinate dialects can be imported into the final component
- Additional data categories in the final component are identified and can if desired be manipulated by conversion routines.



# Structural Integrity

- TBX-Basic fully included sub-set of TBX-Linguist
- Some names changed—mapped to existing names
- Cross-reference involves a structural change
- Some items could be omitted from an input model or display (e.g., /term location/, (term) source)
- Slides 9 & 10 illustrate smooth import

# Tools Issues

- Modify support tools to accommodate new data profile as a superset of TBXBasic
- Lean to rich (if properly mapped) facilitates clean exchange
- Rich to lean – possible tool to convert missing datcats to notes
- Possible if the tool knows about the other dialect

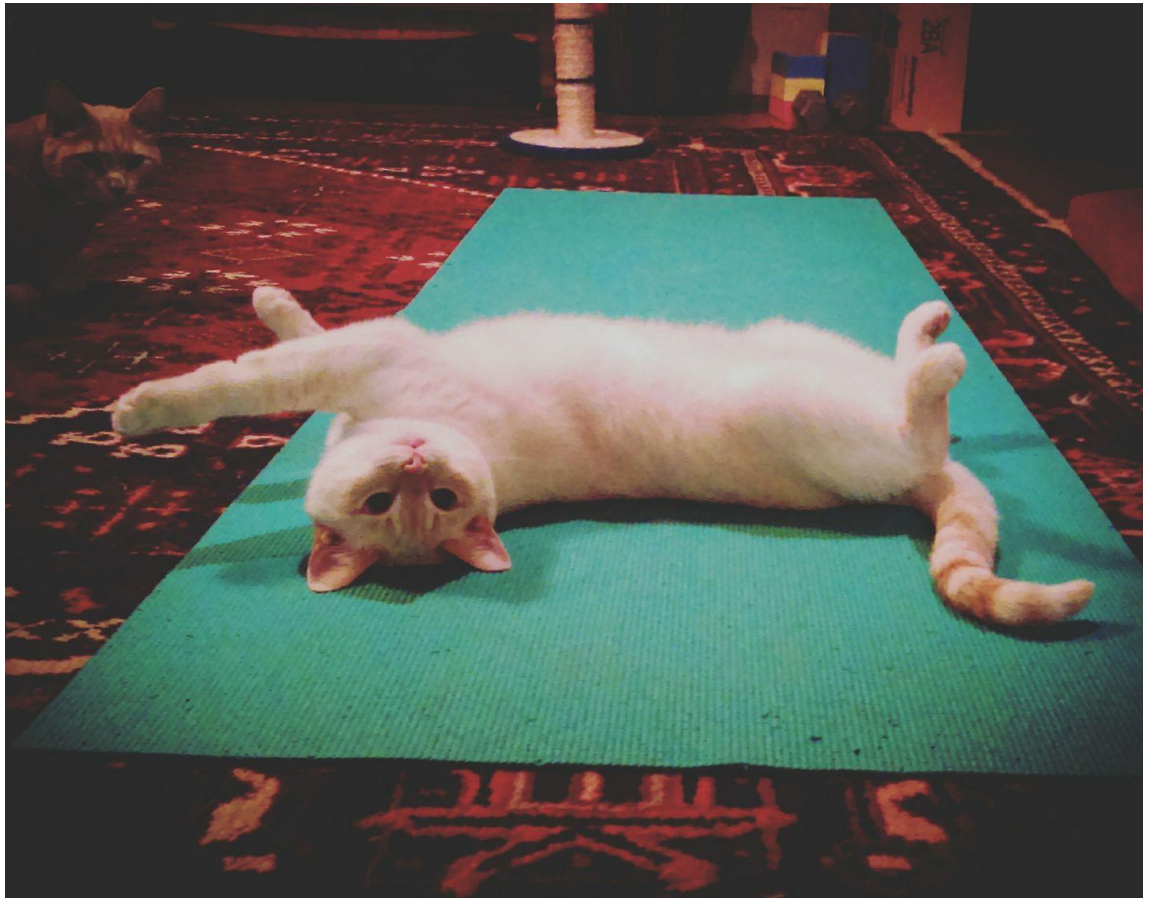
# Additional Issues

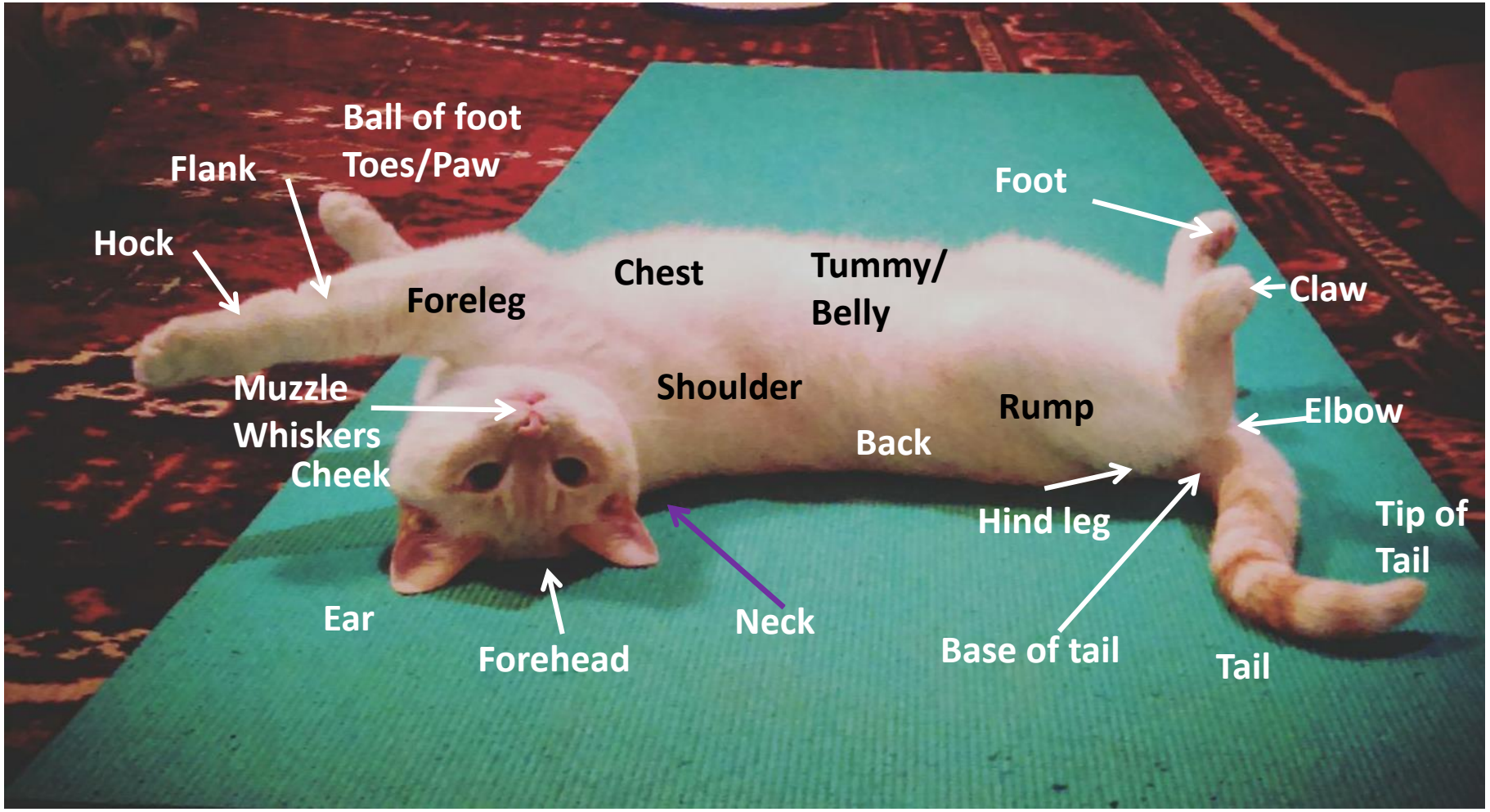
- Concept relations (I-Term, Coreon) and knowledge modeling datcats are currently excluded from the TBX master file.
- Bibliography entries incorporated in termbases are not part of the encoding scheme.
- Coordination with MT component of hybrid CAT/MT?
- Coordination with LMF?

# Contact Information


- Sue Ellen Wright
  swright@kent.edu


**Miss Gina makes her appearance here thanks to the kind permission of her mistress, Jennifer Winer.**




Ball of foot
Toes/Paw
Flank
Foot
Hock
Chest
Tummy/
Belly
Claw
Foreleg
Muzzle
Whiskers
Cheek
Shoulder
Rump
Elbow
Back
Hind leg
Tip of
Tail
Ear
Neck
Forehead
Base of tail
Tail