Unlocking global content with MT and UNL

Mike Dillinger, PhD

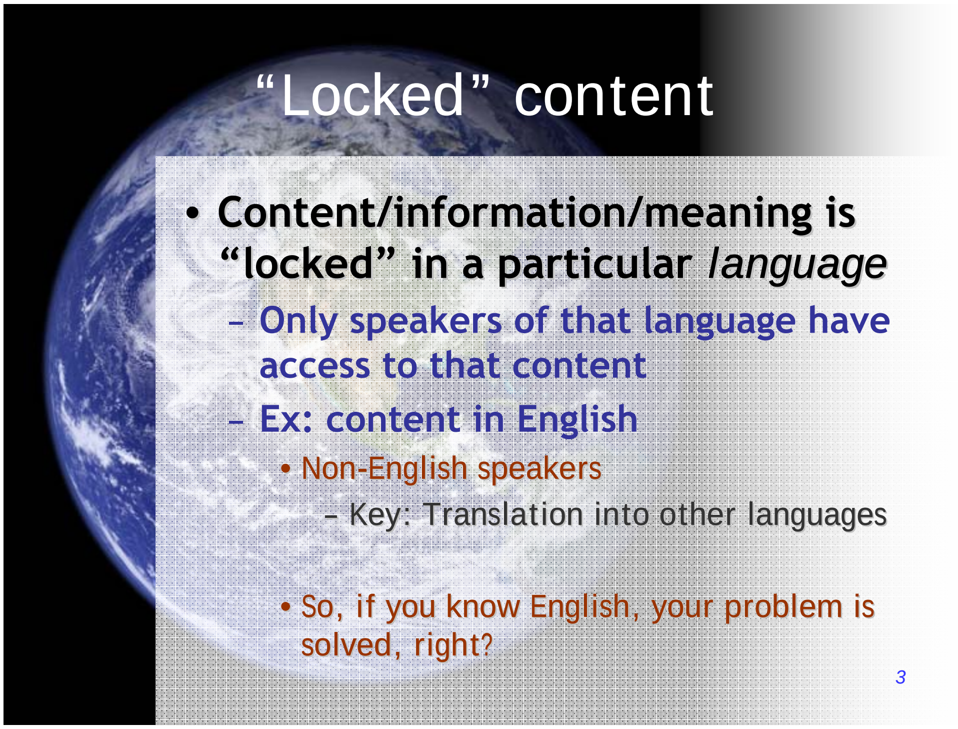
President,

Association for Machine Translation in the Americas

mike@mikedillinger.com

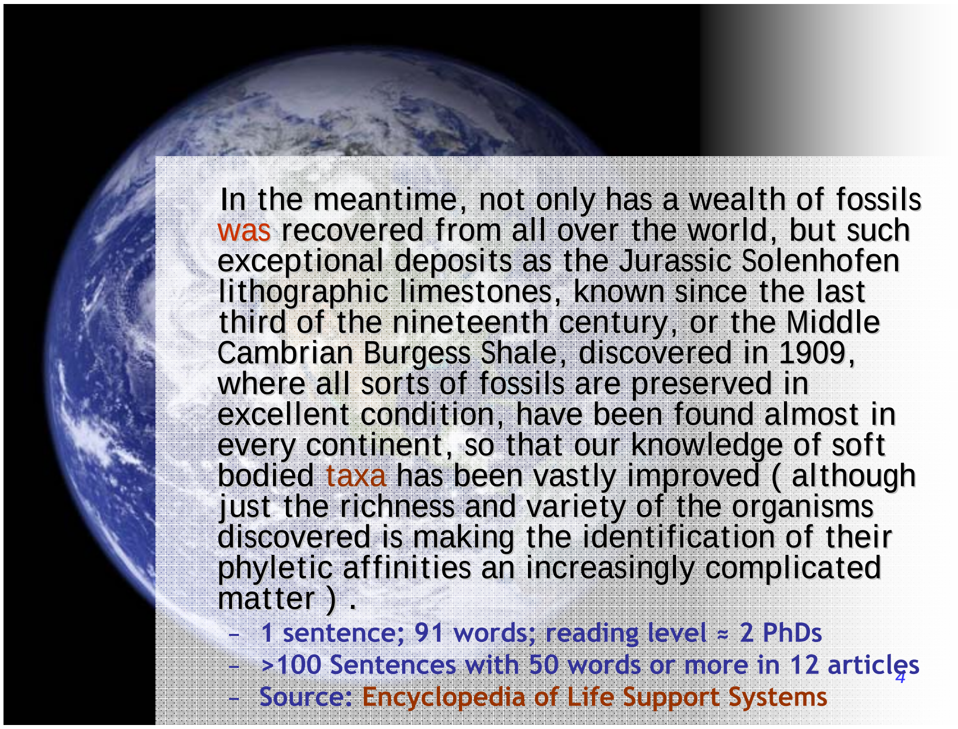
Urgent Warning:

2



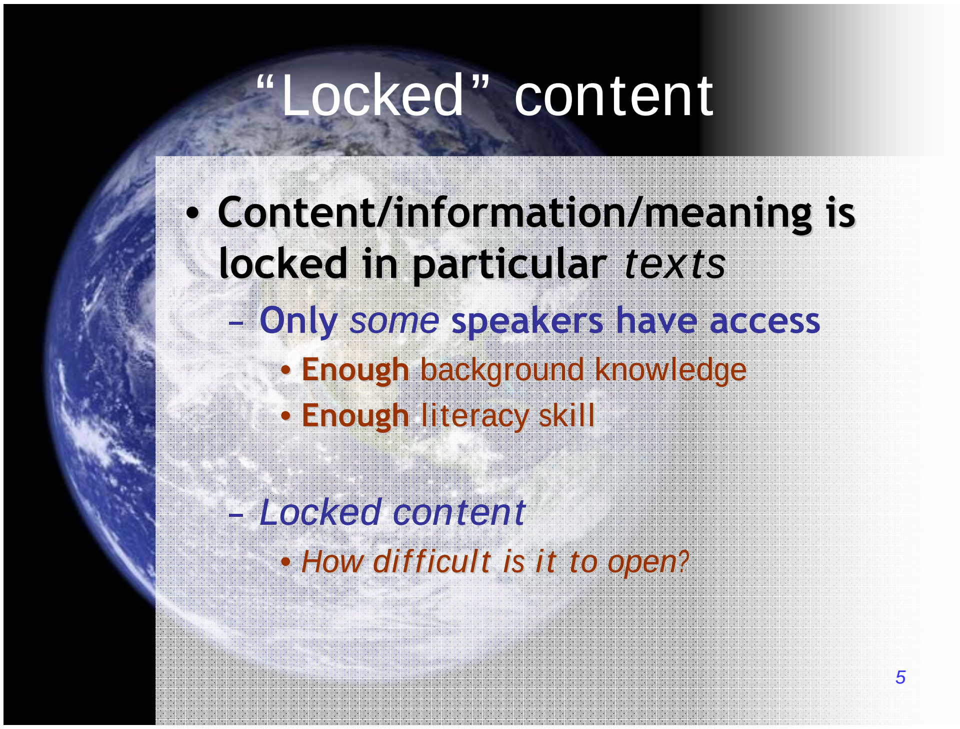
“Locked” content

- Content/information/meaning is “locked” in a particular *language*
  - Only speakers of that language have access to that content
  - Ex: content in English
    - Non-English speakers
      - Key: Translation into other languages
    - So, if you know English, your problem is solved, right?



In the meantime, not only has a wealth of fossils was recovered from all over the world, but such exceptional deposits as the Jurassic Solenhofen lithographic limestones, known since the last third of the nineteenth century, or the Middle Cambrian Burgess Shale, discovered in 1909, where all sorts of fossils are preserved in excellent condition, have been found almost in every continent, so that our knowledge of soft bodied taxa has been vastly improved (although just the richness and variety of the organisms discovered is making the identification of their phyletic affinities an increasingly complicated matter) .

- 1 sentence; 91 words; reading level ≈ 2 PhDs
- >100 Sentences with 50 words or more in 12 articles
- Source: Encyclopedia of Life Support Systems



“Locked” content

- **Content/information/meaning is locked in particular *texts***
  - Only *some* speakers have access
    - **Enough** background knowledge
    - **Enough** literacy skill
  - *Locked content*
    - *How difficult is it to open?*

5

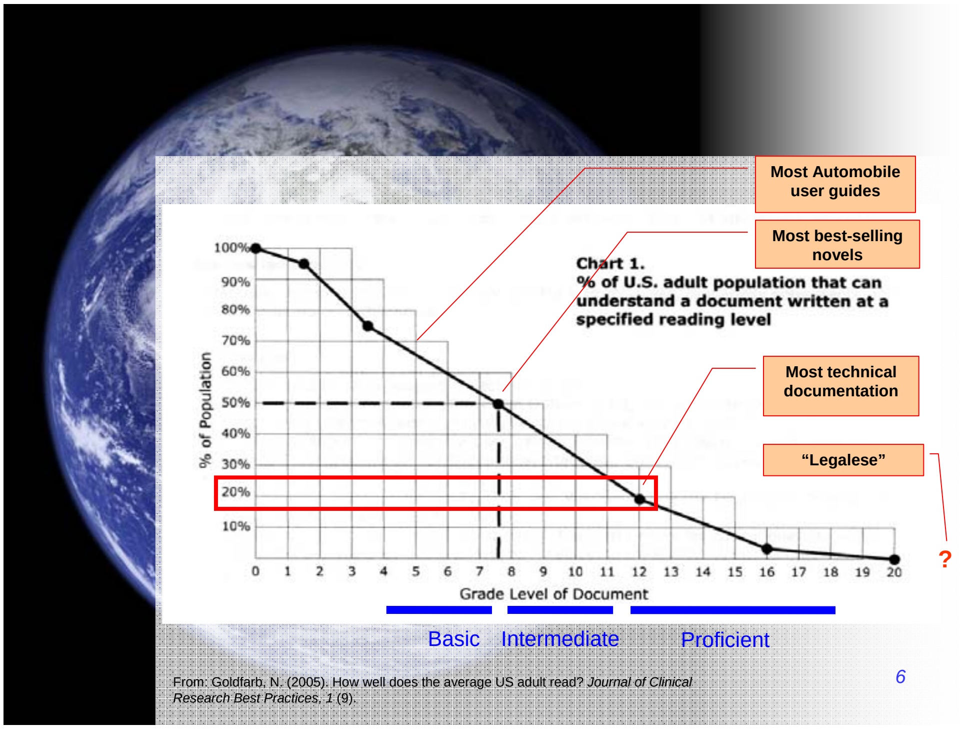


Chart 1.
% of U.S. adult population that can understand a document written at a specified reading level

Most Automobile user guides

Most best-selling novels

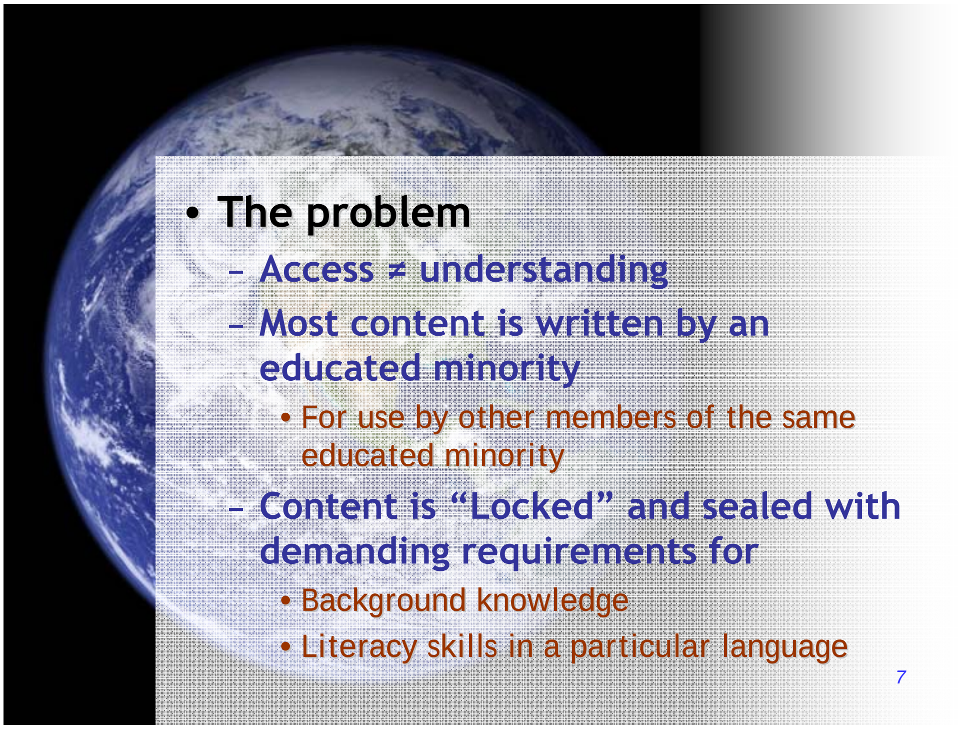
Most technical documentation

“Legalese”

Basic Intermediate Proficient

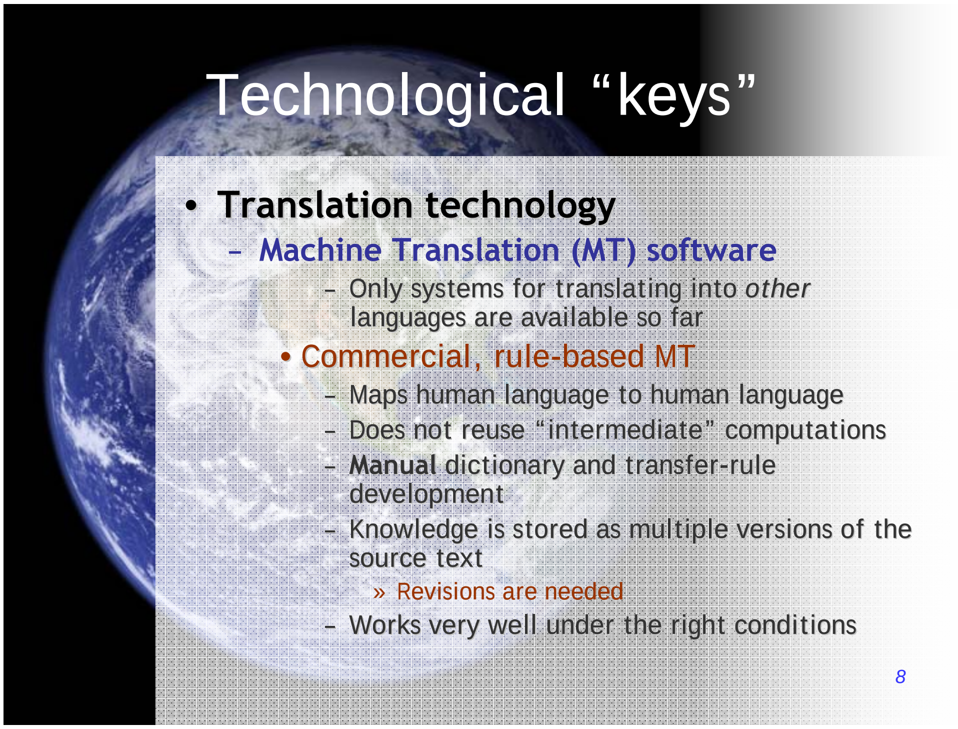
From: Goldfarb, N. (2005). How well does the average US adult read? *Journal of Clinical Research Best Practices, 1* (9).

6



- The problem
  - Access ≠ understanding
  - Most content is written by an educated minority
    - For use by other members of the same educated minority
  - Content is “Locked” and sealed with demanding requirements for
    - Background knowledge
    - Literacy skills in a particular language

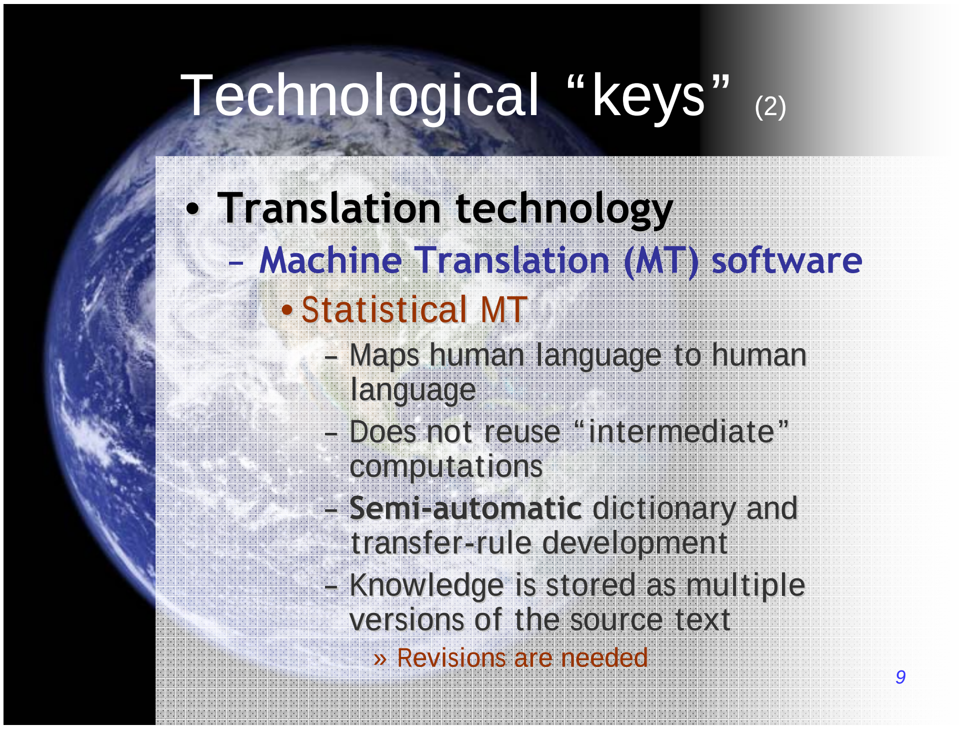
7



Technological “keys”

- Translation technology
  - Machine Translation (MT) software
    - Only systems for translating into *other* languages are available so far
    - Commercial, rule-based MT
      - Maps human language to human language
      - Does not reuse “intermediate” computations
      - **Manual** dictionary and transfer-rule development
      - Knowledge is stored as multiple versions of the source text
        - » Revisions are needed
      - Works very well under the right conditions

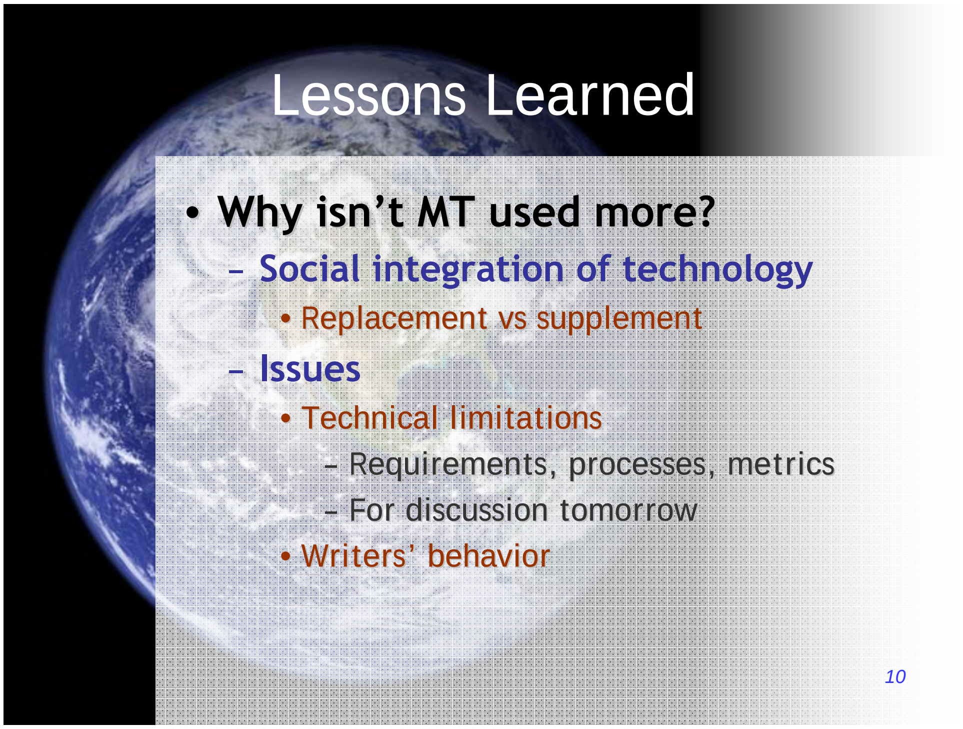
8



Technological "keys" (2)

- Translation technology
  - Machine Translation (MT) software
    - Statistical MT
      - Maps human language to human language
      - Does not reuse "intermediate" computations
      - **Semi-automatic** dictionary and transfer-rule development
      - Knowledge is stored as multiple versions of the source text
        - » Revisions are needed

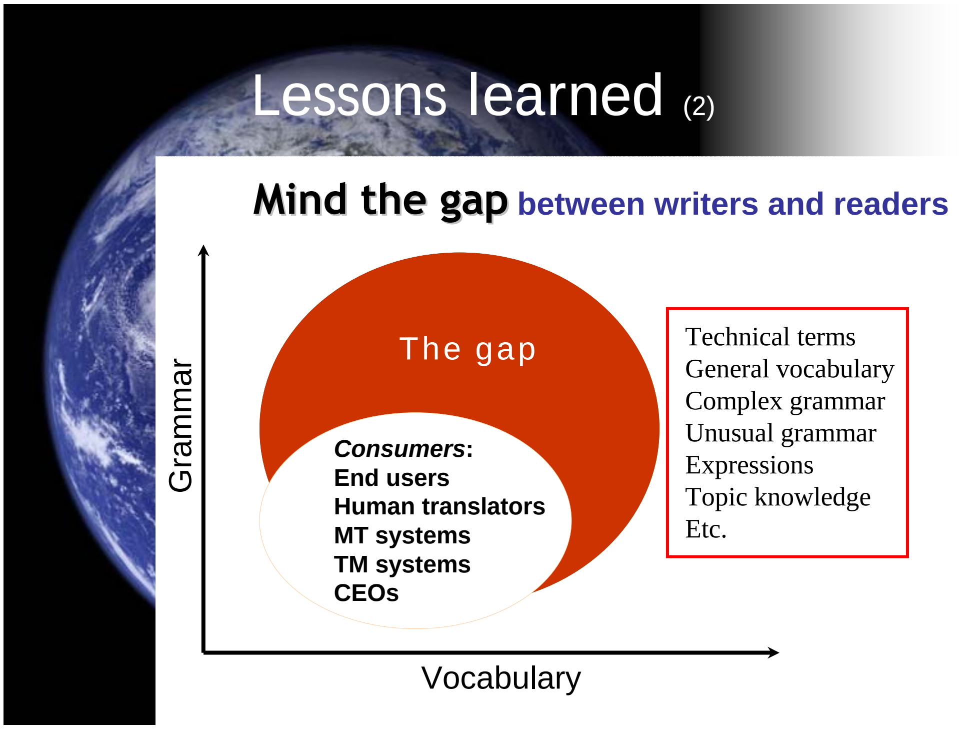
9



Lessons Learned

- Why isn't MT used more?
  - Social integration of technology
    - Replacement vs supplement
  - Issues
    - Technical limitations
      - Requirements, processes, metrics
      - For discussion tomorrow
    - Writers' behavior

10



Lessons learned (2)

Mind the gap between writers and readers

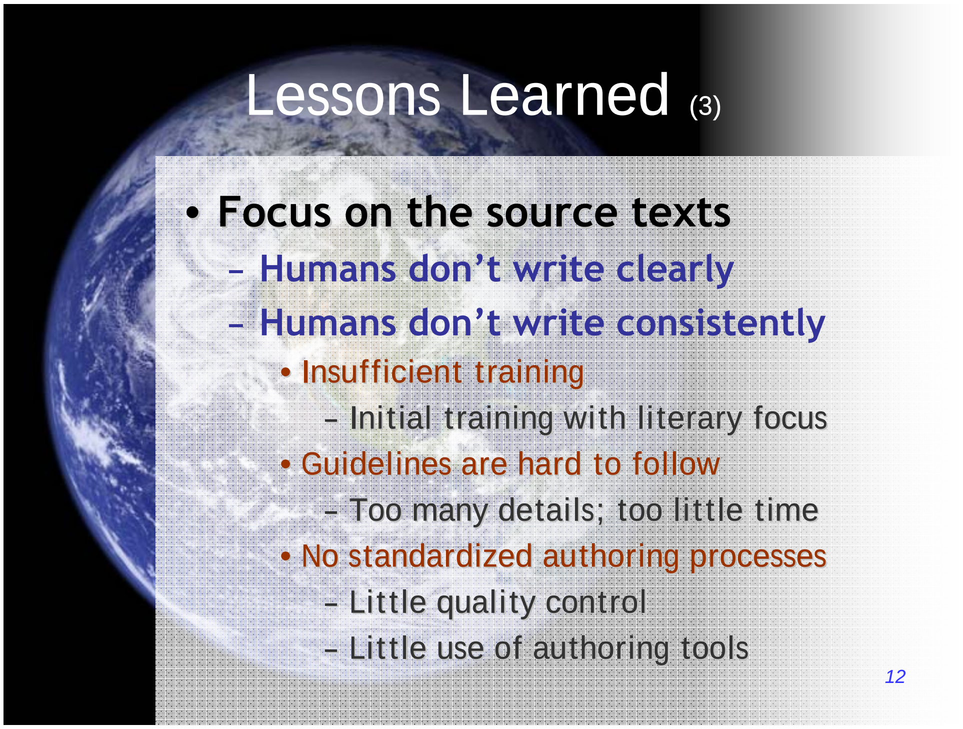
Grammar

The gap

***Consumers*:**
End users
Human translators
MT systems
TM systems
CEOs

Vocabulary

Technical terms
General vocabulary
Complex grammar
Unusual grammar
Expressions
Topic knowledge
Etc.



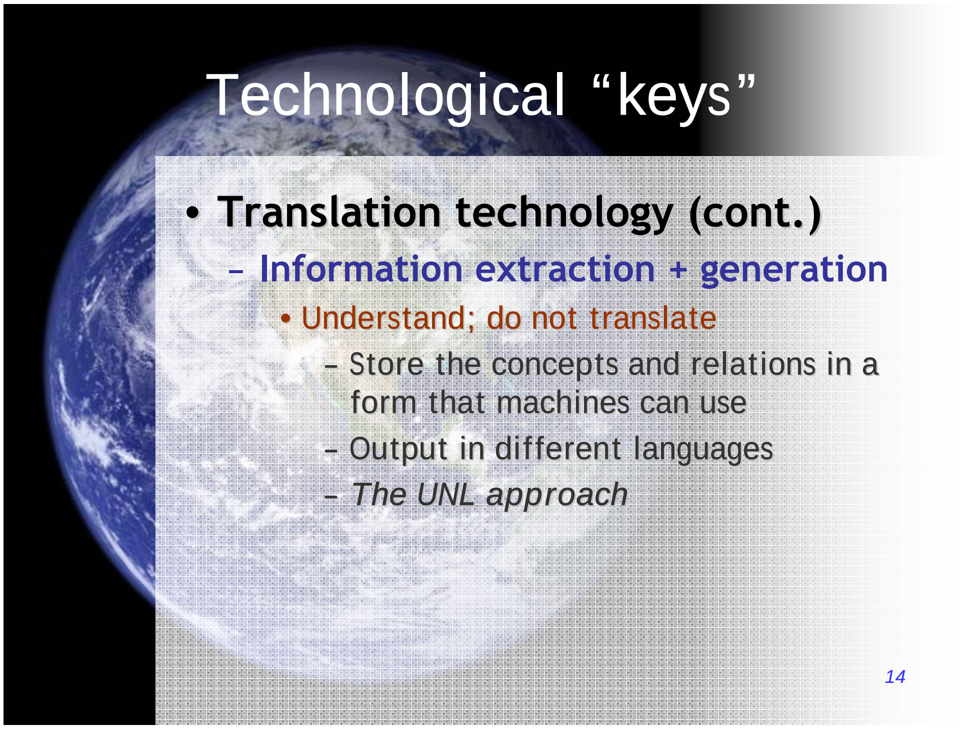
Lessons Learned (3)

- Focus on the source texts
  - Humans don’t write clearly
  - Humans don’t write consistently
    - Insufficient training
      - Initial training with literary focus
    - Guidelines are hard to follow
      - Too many details; too little time
    - No standardized authoring processes
      - Little quality control
      - Little use of authoring tools

12

Lessons Learned (4)

- **Focus on the source texts**
  - **Action item:**
    - Adapt source texts to the limitations of translation technology
      - Revise source texts
        - » More readable for humans; more readable for machines
      - Adapting technology to human variability is not cost effective in the short term
        - » We can’t wait for perfect technology



Technological “keys”

- **Translation technology (cont.)**
  - **Information extraction + generation**
    - Understand; do not translate
      - Store the concepts and relations in a form that machines can use
      - Output in different languages
      - *The UNL approach*

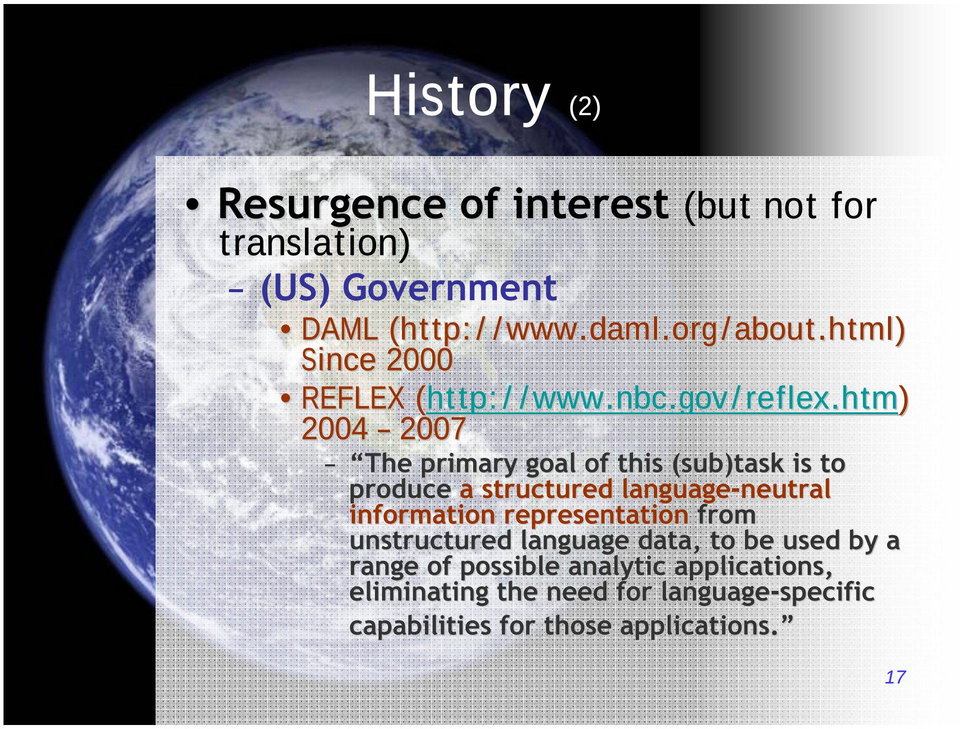
- Global access to content
  - The UNL approach is formulated to meet the widest range of content-access needs
    - Within the same language: generate more readable or more technical texts
    - Across languages
  - Next-generation knowledge processing architectures will have most or all of the characteristics of UNL
    - E.g., Semantic Web, ontologies, etc.

15

History

- From text to knowledge
    - "Automated understanding", "information extraction", etc.
  - Started in the early 1970s
    - Schank, Bobrow, Kintsch, Frederiksen, etc.
  - Loss of interest around 1990
    - Too much manual work need for applications
    - Novelty of statistical methods
  - Keep the torch burning
    - TREC (http://trec.nist.gov/)
      - Since 1992 (from the Tipster project)
    - UNL
    - TANKA (University of Ottawa)

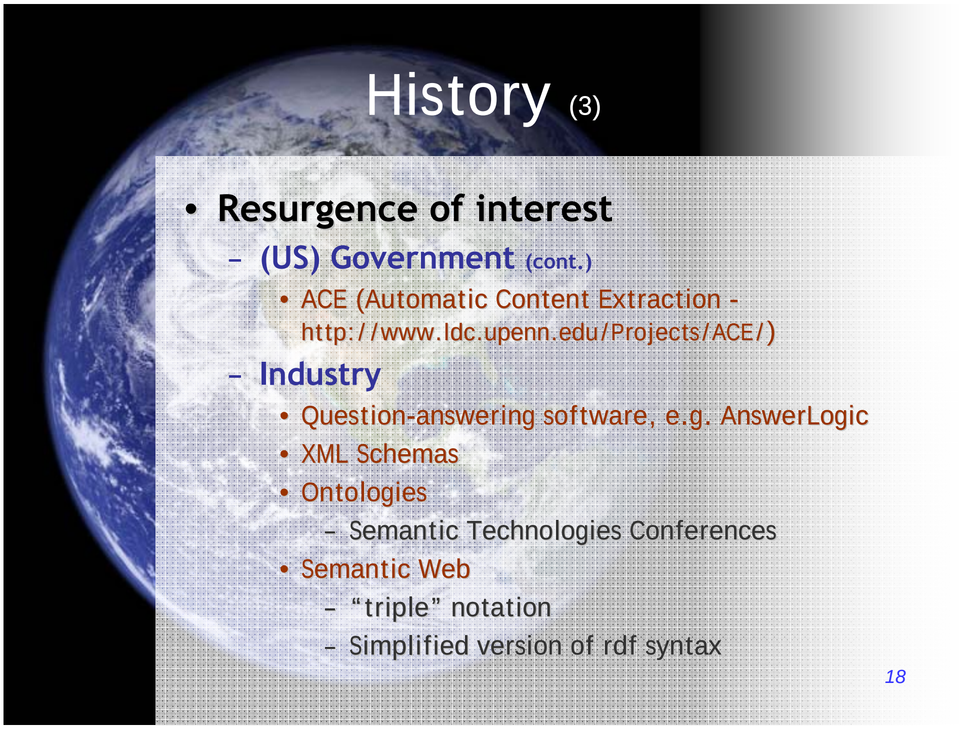
16



History (2)

- **Resurgence of interest** (but not for translation)
  - **(US) Government**
    - DAML (http://www.daml.org/about.html) Since 2000
    - REFLEX (http://www.nbc.gov/reflex.htm) 2004 – 2007
      - "The primary goal of this (sub)task is to produce a structured language-neutral information representation from unstructured language data, to be used by a range of possible analytic applications, eliminating the need for language-specific capabilities for those applications."

17



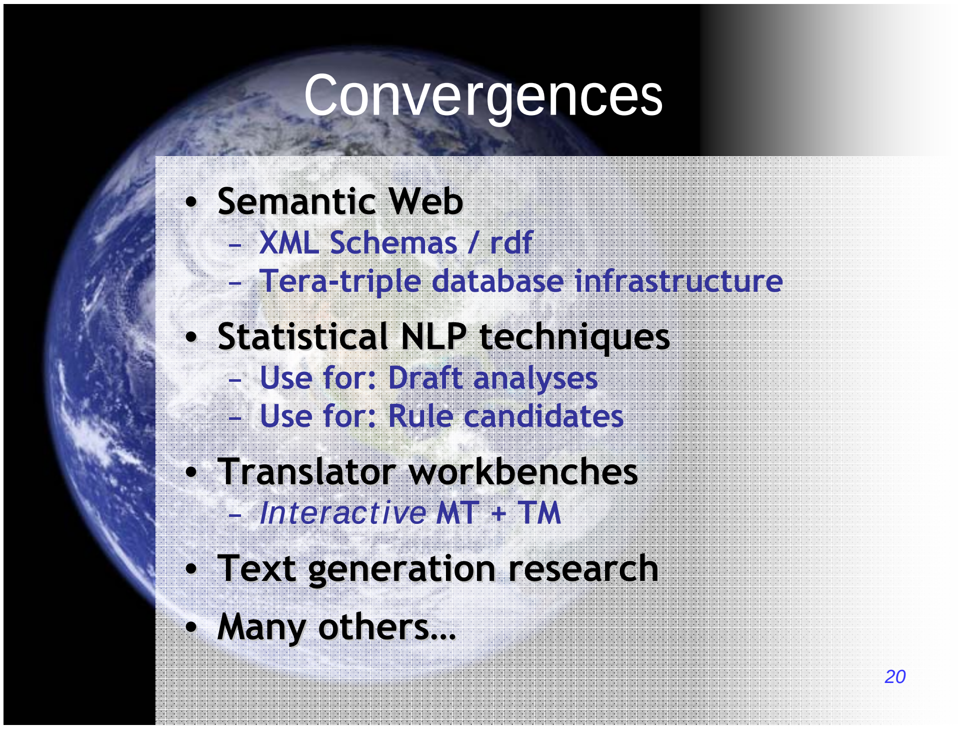
History (3)

- **Resurgence of interest**
  - **(US) Government** (cont.)
    - ACE (Automatic Content Extraction - http://www.ldc.upenn.edu/Projects/ACE/)
  - **Industry**
    - Question-answering software, e.g. AnswerLogic
    - XML Schemas
    - Ontologies
      - Semantic Technologies Conferences
    - Semantic Web
      - "triple" notation
      - Simplified version of rdf syntax

18

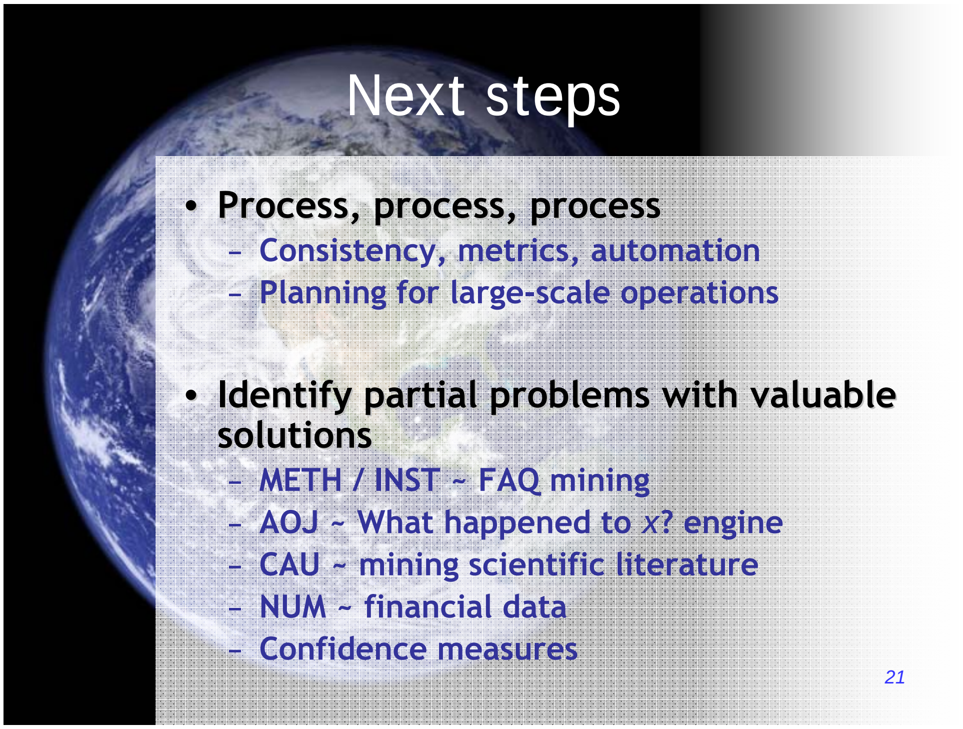
History (4)

- Resurgence of interest
  - Academic research
    - Semantic bootstrapping (Riloff, 1998); Text Mining (Hearst, 1999)
    - FrameNet (www.icsi.berkeley.edu/~framenet/ - since 1999)
    - D. Gildea & D. Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics,* 28:3, 245-288.
    - Semantic Role Mapping “bake-offs”
      - Since 2004: CoNLL, SENSEVAL, etc.
    - Helbig, Hermann (2006). *Knowledge Representation and the Semantics of Natural Language*. Berlin: Springer. [MultiNet]



Convergences

- Semantic Web
  - XML Schemas / rdf
  - Tera-triple database infrastructure
- Statistical NLP techniques
  - Use for: Draft analyses
  - Use for: Rule candidates
- Translator workbenches
  - *Interactive* MT + TM
- Text generation research
- Many others…



Next steps

- Process, process, process
  - Consistency, metrics, automation
  - Planning for large-scale operations
- Identify partial problems with valuable solutions
  - METH / INST ~ FAQ mining
  - AOJ ~ What happened to *x*? engine
  - CAU ~ mining scientific literature
  - NUM ~ financial data
  - Confidence measures

21

UNL …

for Universal Access to Knowledge