Language Archives and Linguistic Anchoring of Digital Archives

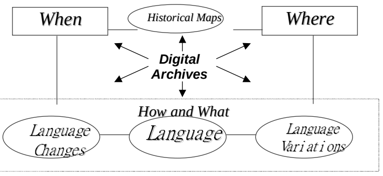
Chu-Ren Huang
Institute of Linguistics, Academia Sinica

LSA Symposium:
The Open Language Archives Community
4 January 2002

Linguistic Anchoring of Digital Archives

- Language Archives serve communities beyond linguists
- Linguistic description and interpretation underlies any digital archive items
- In digital archives, each knowledge item should be temporally, geographically, and ***linguistically*** anchored.

Language and Digital Archives



Digital Archives are Linguistically Anchored

- **Archives are anchored with Lexical KnowledgeBase (LKB)**

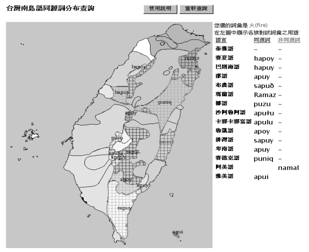
-because LKB as collection of lexical types instantiated in archives uniquely defines each archive

-And each lexical item is the conceptual atom projecting knowledge from archive to archive

From Linguistic Anchor to Knowledge Projection

- Synergy of language archives anchored by lexical forms and supported by LKB generates new knowledge
- Extension of linguistic anchoring based on LKB to all types of digital archives will lead to even more creative synergy

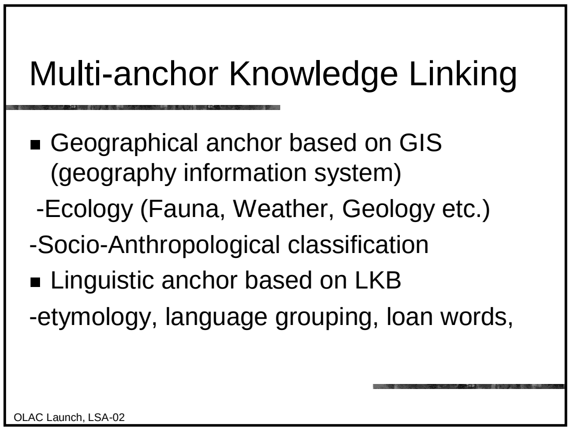
Where & What: Language Atlas



台灣南島語同源詞分布查詢

您選的詞彙是 火(fire)
在左圖中顯示各族對該詞彙之用語:

語言	同源詞	非同源詞
泰雅語	–	–
賽夏語	hapoy	–
巴則海語	hapuy	–
邵語	apuy	–
布農語	sapuð	–
噶蘭語	Ramaz	–
鄒語	puzu	–
沙阿魯阿語	apuɫu	–
卡那卡那富語	apulu	–
魯凱語	apoy	–
排灣語	sapuy	–
卑南語	apuy	–
賽德克語	puniq	–
阿美語		namal
雅美語	apui	



Multi-anchor Knowledge Linking

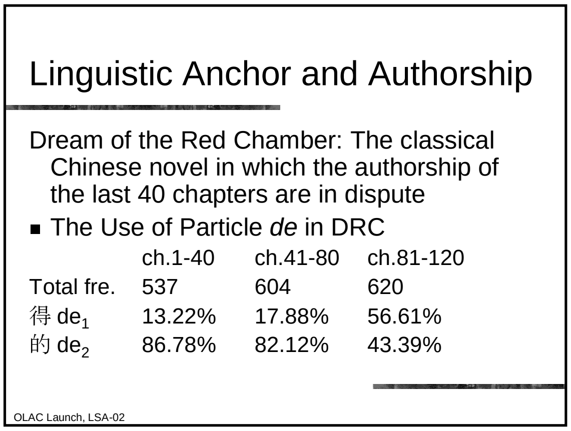
- Geographical anchor based on GIS (geography information system)

-Ecology (Fauna, Weather, Geology etc.)

-Socio-Anthropological classification

- Linguistic anchor based on LKB

-etymology, language grouping, loan words,

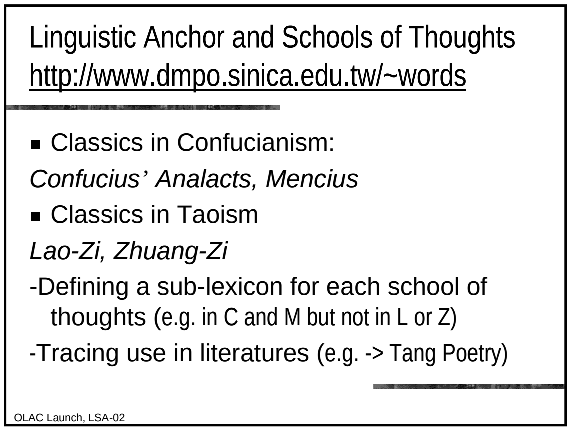


Linguistic Anchor and Authorship

Dream of the Red Chamber: The classical Chinese novel in which the authorship of the last 40 chapters are in dispute

- The Use of Particle *de* in DRC

	ch.1-40	ch.41-80	ch.81-120
Total fre.	537	604	620
得 de_1	13.22%	17.88%	56.61%
的 de_2	86.78%	82.12%	43.39%



Linguistic Anchor and Schools of Thoughts http://www.dmpo.sinica.edu.tw/~words

- Classics in Confucianism:

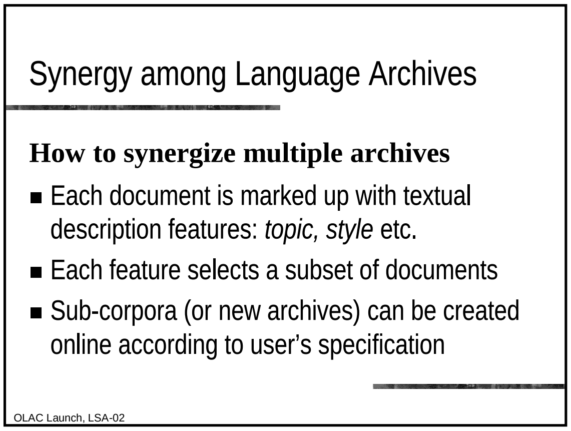
Confucius' Analacts, Mencius

- Classics in Taoism

Lao-Zi, Zhuang-Zi

-Defining a sub-lexicon for each school of thoughts (e.g. in C and M but not in L or Z)

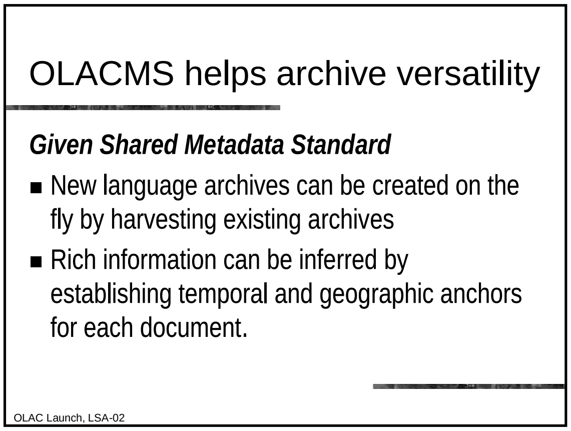
-Tracing use in literatures (e.g. -> Tang Poetry)



Synergy among Language Archives

How to synergize multiple archives

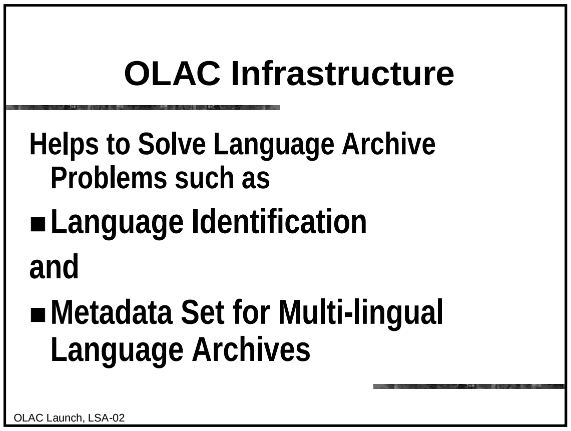
- Each document is marked up with textual description features: *topic, style* etc.
- Each feature selects a subset of documents
- Sub-corpora (or new archives) can be created online according to user's specification



OLACMS helps archive versatility

Given Shared Metadata Standard

- New language archives can be created on the fly by harvesting existing archives
- Rich information can be inferred by establishing temporal and geographic anchors for each document.



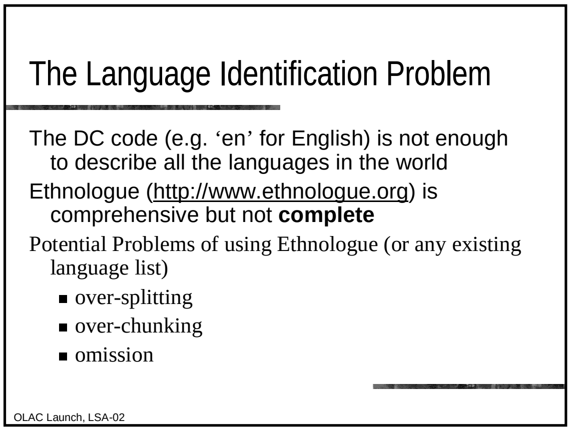
OLAC Infrastructure

Helps to Solve Language Archive Problems such as

- **Language Identification**

and

- **Metadata Set for Multi-lingual Language Archives**



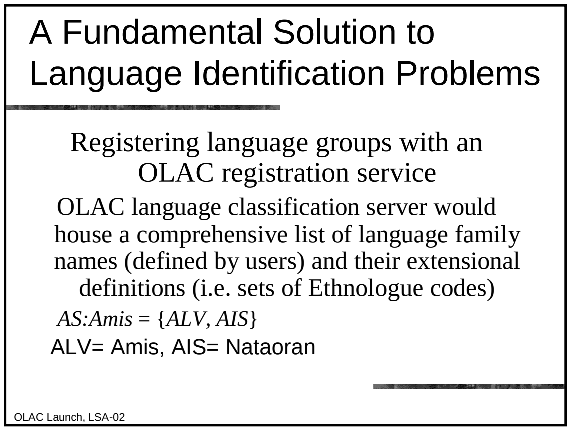
The Language Identification Problem

The DC code (e.g. ‘en’ for English) is not enough to describe all the languages in the world

Ethnologue (http://www.ethnologue.org) is comprehensive but not **complete**

Potential Problems of using Ethnologue (or any existing language list)

- over-splitting
- over-chunking
- omission



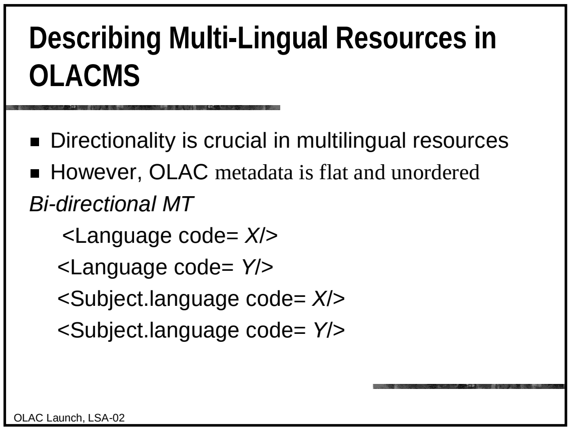
A Fundamental Solution to Language Identification Problems

Registering language groups with an OLAC registration service

OLAC language classification server would house a comprehensive list of language family names (defined by users) and their extensional definitions (i.e. sets of Ethnologue codes)

AS:Amis = {*ALV*, *AIS*}

ALV= Amis, AIS= Nataoran



Describing Multi-Lingual Resources in OLACMS

- Directionality is crucial in multilingual resources
- However, OLAC metadata is flat and unordered

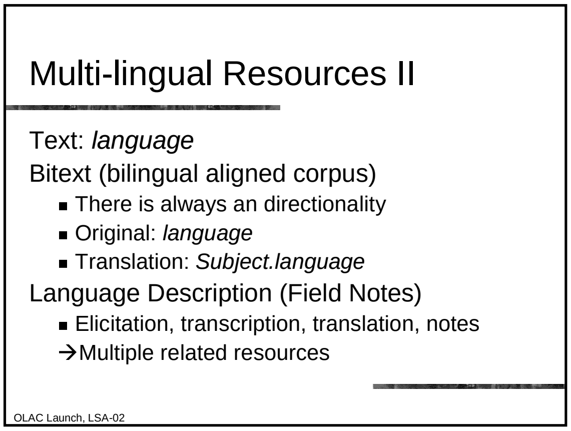
Bi-directional MT

<Language code= *X*/>

<Language code= *Y*/>

<Subject.language code= *X*/>

<Subject.language code= *Y*/>



Multi-lingual Resources II

Text: *language*

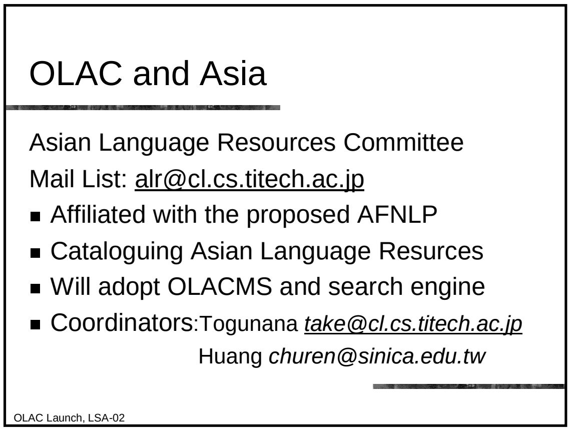
Bitext (bilingual aligned corpus)

- There is always an directionality
- Original: *language*
- Translation: *Subject.language*

Language Description (Field Notes)

- Elicitation, transcription, translation, notes

→Multiple related resources

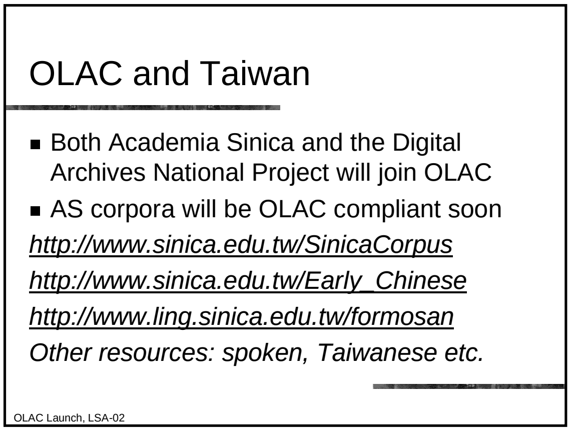


OLAC and Asia

Asian Language Resources Committee

Mail List: alr@cl.cs.titech.ac.jp

- Affiliated with the proposed AFNLP
- Cataloguing Asian Language Resures
- Will adopt OLACMS and search engine
- Coordinators:Togunana *take@cl.cs.titech.ac.jp*
  Huang *churen@sinica.edu.tw*



OLAC and Taiwan

- Both Academia Sinica and the Digital Archives National Project will join OLAC
- AS corpora will be OLAC compliant soon

http://www.sinica.edu.tw/SinicaCorpus

http://www.sinica.edu.tw/Early_Chinese

http://www.ling.sinica.edu.tw/formosan

Other resources: spoken, Taiwanese etc.