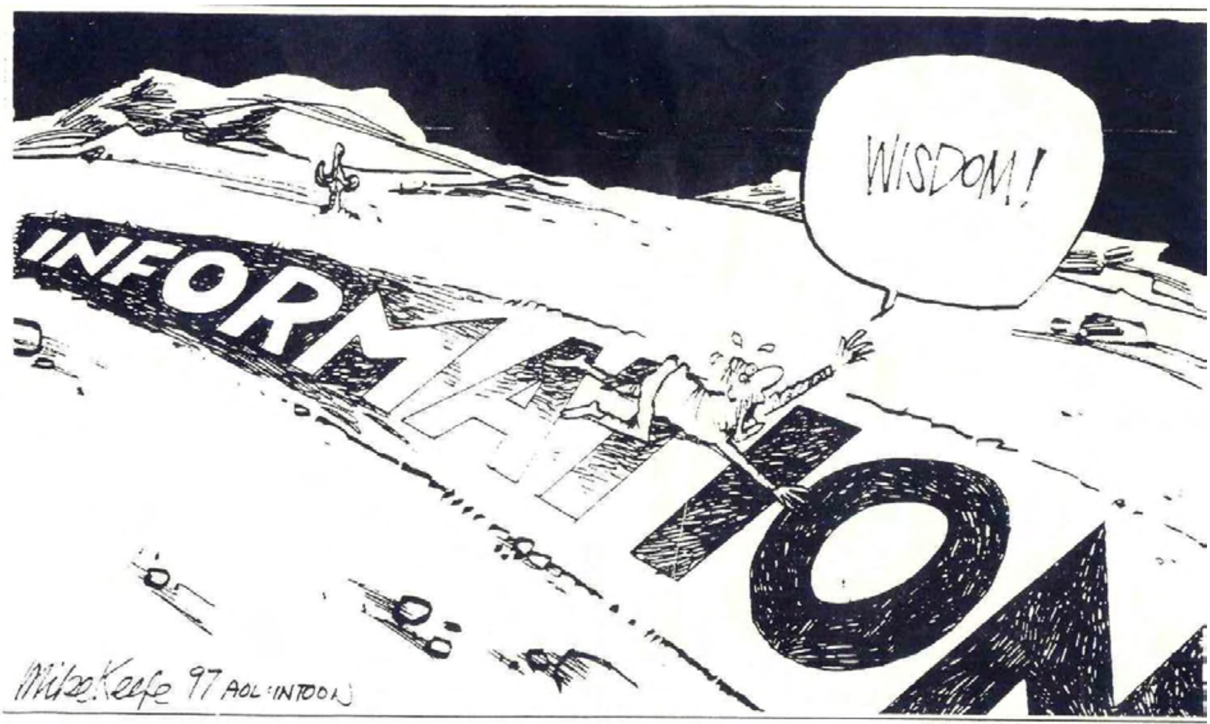


By Mike Keefe, The Denver Post, for USA TODAY

RESEARCH INSTITUTE FOR LINGUISTICS
HUNGARIAN ACADEMY OF SCIENCES

MAGYAR TUDOMÁNYOS AKADÉMIA 1825

E-Reference Database for the Humanities: ERIH-Online

István Kenesei
RIL HAS

Why collect references?

- Measure impact of research for a multitude of purposes
- Long introduced and utilised in sciences
- (Emerging field of scientometrics, complex calculations of impact factor: e.g., ratio of citations and references per year of a journal)
- Humanities lagging behind - **Why?**

First and foremost: fragmentation

Minimally two-way fragmentation:

1. Less theory-oriented fields with several schools, trends hardly in communication with each other
2. In spite of common and/or overarching interests, fields are partitioned by national or local concerns and, most importantly by the simple communicative device of **language**.

Publications are in several languages and practitioners often make references to works written in various languages.

The sheer fact of a multilingual target may scare away any light-hearted research appraiser.

Related issue: Should we keep up publishing in native/national languages?

My conviction: Necessary for the good of future generations

Secondly:

What's available now? (Apart from manual search)

A) Scholar Google:

Problems:

- only from internet, whether actually published or found on a URL
- duplicates not screened
- same work under different/several entries

→ haphazard, mixture of real and faulty references, including references to works by authors in volumes edited by themselves;

Should be combined with Google Books
– but that serves a different purpose

B) Harzing's Publish or Perish

based on Scholar Google, same predicament, though flurry of calculations from h-index to e-index

Difficulties:

- identical names (e.g., Brody)
- areas/disciplines not kept clearly apart
- manual screening necessary
- gives no references (unlike Scholar Google), only statistics (no wonder: that was the purpose!)

C) Thomson Institute's (ISI) Web of Science

- Definitely *the* information source in the field
- Its achievements are difficult to surpass, though not impossible
- Has a long history: it has collected data from 1981
- Has a well-designed cumulative database
- Applies a pool of highly developed analytic tools evaluating and classifying journals, researchers, institutions, countries, etc.
- Its evaluations are based on a meticulous research into references to journals, continuously updated.
- A business enterprise with a world-wide network
- Has built up its reputation on strict methods and reliable analyses

Problems with ISI

- Limited input: only from journals screened
- Limited number of references: again only from same journals
- Haphazard selection of books
- No references from books
- OK as source of ‘official’ IF – but no good as source of cumulative impact, i.e., list of all references to a paper

Where ERIH can fare better than ISI:

Numbers:

In linguistics, ISI currently has 104 major titles out of a total of 458 titles, ERIH has 586

Languages:

- ISI’s principle of language selection:

> “English is the universal language of science at this time in history. It is for this reason that Thomson Scientific focuses on journals that publish full text in English or at very least, their bibliographic information in English. There are many journals covered in Web of Science that publish only their bibliographic information in English with full text in another language. ”

- ISI has journals in 6 languages:

  English, French, German, Spanish, Italian and Croatian (!)
- ERIH covers (almost) all European languages.

ISI's fields and coverage

Psychology / Psychiatry	Psychiatry Psychology
Social Sciences, General	Anthropology Communication Environmental Studies, Geography & Development Library & Information Science Political Science & Public Administration Public Health & Health Care Science Rehabilitation Social Work & Social Policy Sociology & Social Sciences
Arts & Humanities Categories are not included in the Standard Indicators database	Archeology Art & Architecture Classical Studies General History Language & Linguistics Literature Performing Arts Philosophy Religion & Theology

ERIH’s fields and coverage

- Anthropology
- Archaeology
- Art and Art History
- Classical Studies
- Gender Studies
- History
- History and Philosophy of Science
- Linguistics
- Literature
- Musicology
- Oriental and African Studies
- Pedagogical and Educational Research
- Philosophy
- Psychology
- Religious Studies and Theology

Where to go from here?

Build e-database from references of articles in journals listed and classified in ERIH

Input:

- Language technology capable of handling multilingual publications
- Multiple scripts possible (NB. ISI is restricted to the Roman alphabet)
- Access (only) to reference sections of journals
- Clearance of property rights

Output:

- Easily accessible web-searchable reference database

WORK PLAN

I. **Feasibility study**

a) Exploring the resource requirements

- accessibility of journals (online/offline/print only)
- estimate of the number of references (based on the number of journals/issues and average number of references)
- estimates based on a tally of the references found in one volume each from a random sample of 26 journals suggest that the database will yield c. 120 million database records

b) clearing IPR issues

c) working out a detailed project plan

Labour: 6-9 man-months

II. Data acquisition

a) Locate and contact publishers
b) Negotiate licences
c) Acquire and archive source data
Labour: 12-15 man-months

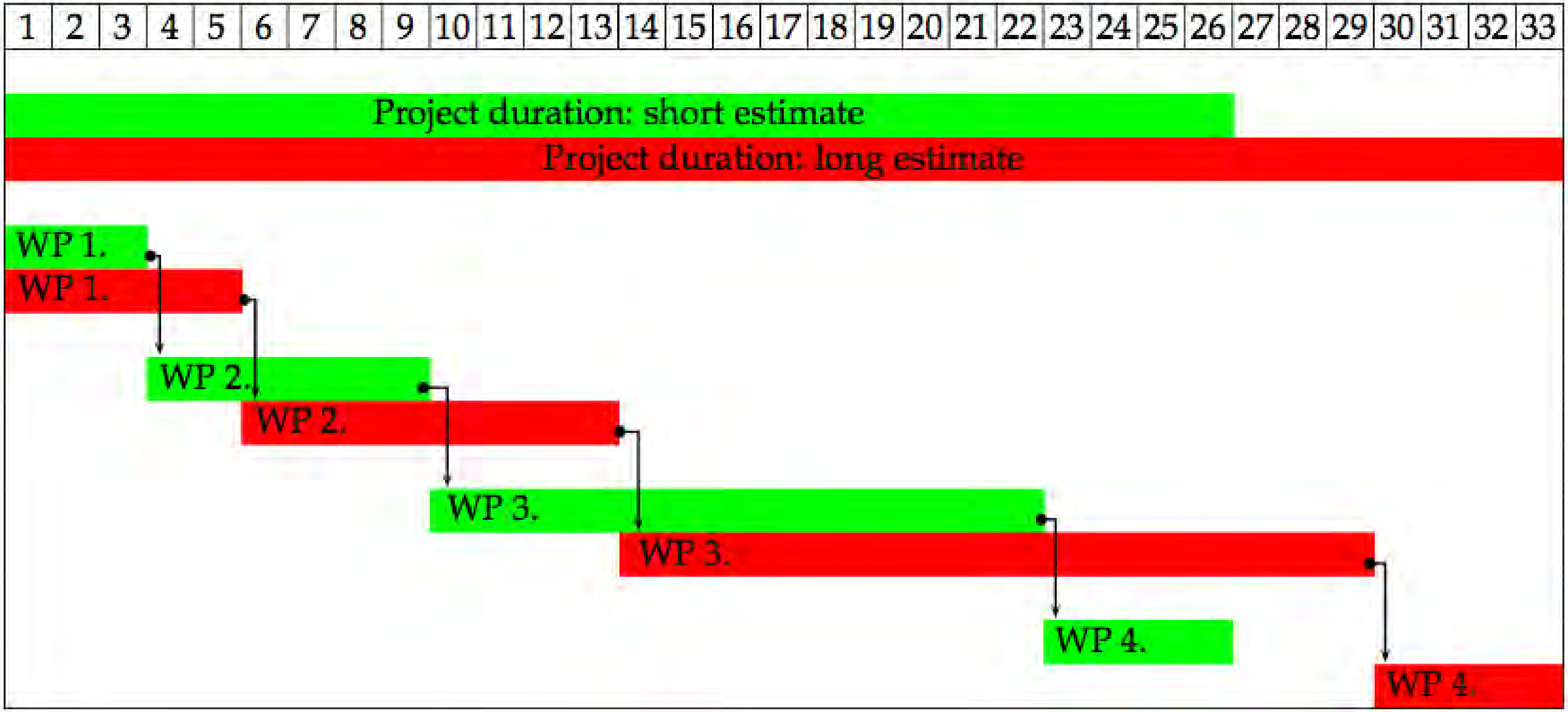
III. Database design and implementation

a) Design of the database
b) preprocessing of the data
c) populating the database with data
d) testing and putting the database into operation
Labour: 80-100 man-months

IV. Web access

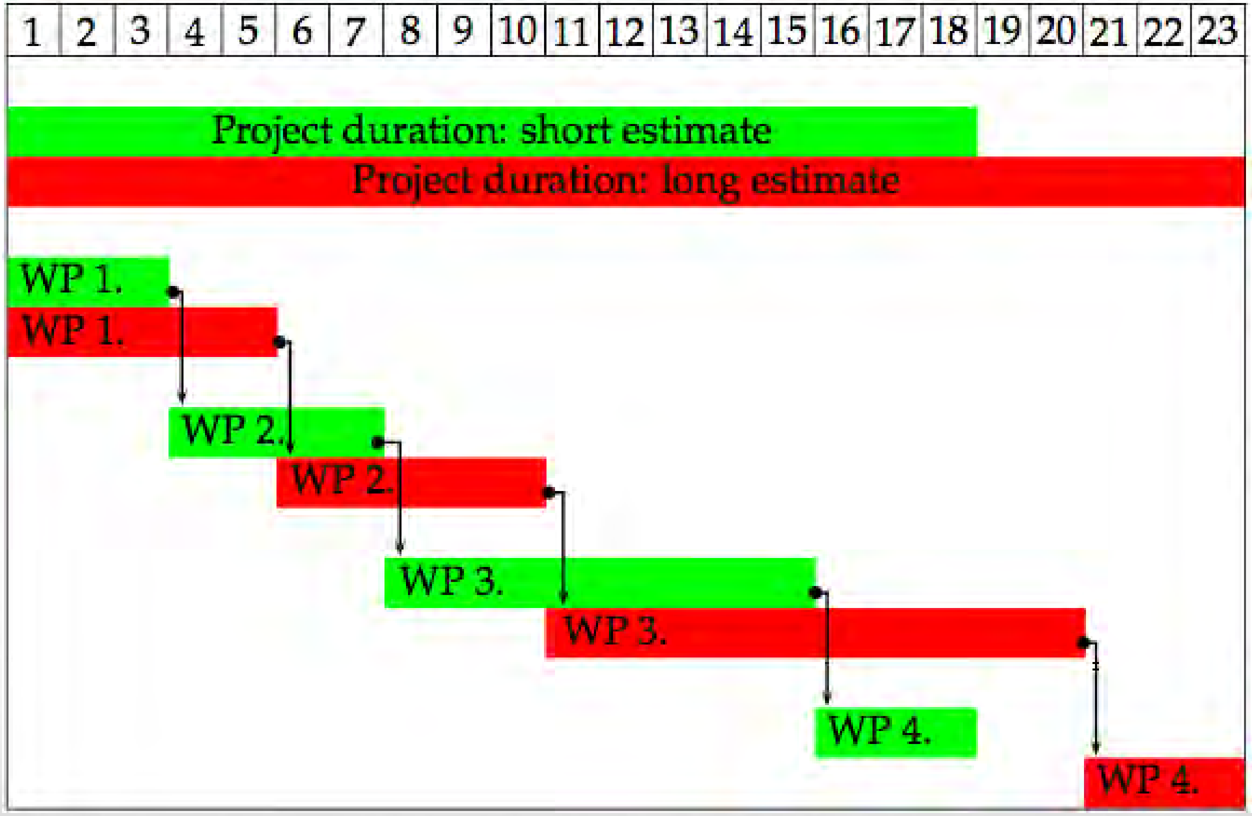
a) database interface
b) web access interface
Labour: 12 man-months

Workpackage schedule



Scheduling of workpackages: lower number of personnel

Workpackage schedule



Scheduling of workpackages: higher number of personnel

Further avenues:

- Academics-based (peer-reviewed) classification of journals to be complemented by their weight-based analysis of cross-referencing (cf. ISI’s practice)

- Including monographs (a crucial factor in humanities) will add to ERIH’s edge over ISI

- Question: Can ISI be amalgamated with ERIH-Online?
  – No: Even class A references underrepresented in ISI

- Additional advantage: gives prestige to publications in native/national languages

- Serves as example for other multilingual regions

- In foreseeable future: ERIH-Online can easily become self-supporting from publishers’ ads and other sponsors

Possible pilot study

- Project for a complete reference data-base of journals in social sciences & humanities published in Hungary
- About 200 journal numbers, 8 million records, based on 25 journals
- But: limited number of languages (Hungarian, English)
- Time span: 24 months
- Possible sponsors: HAS, Rectors' Conference, OTKA

Acknowledgements:

- Tamás Váradi – for collaboration on project proposal
- Judit Kuti – for help to produce the slides

Thank you for your attention!