



evaluation of digital libraries: an overview

Tefko Saracevic, Ph.D.

School of Communication, Information and
Library Studies

Rutgers University

<http://www.scils.rutgers.edu/~tefko>



“Evaluating digital libraries is a bit like judging how successful is a marriage”

(Marchionini, 2000)



digital libraries

- since emergence in early/mid 1990's
 - many institutions & fields got involved
 - great many practical developments
 - many research efforts & programs globally
 - large expenditures in research & practice
 - applications & use growing exponentially
- everything about digital libraries is explosive
- **except evaluation**
 - relatively small, even neglected area



literature reports on DL evaluation

- two distinct types:
 - meta or “about” literature
 - suggest approaches, models, concepts
 - discussed evaluation
 - object or “on” literature
 - actual evaluations, contains data
 - data could be hard or soft
- meta literature much larger
 - parallel with IR evaluation literature in 1960's & early 70's



objective & corpus

- to synthesize **object** literature only
- selection criteria:
 1. directly address a DL entity or a DL process
 2. contain data in whatever form
- some 80 reports selected
- estimate: no more than 100 or so evaluation reports exist totally



boundaries

- difficult to establish, apply
 - particularly as to process – e.g.
 - crossing into IR: where does IR evaluation stop & DL evaluation start?
 - or any technology evaluation?
 - or evaluation of web resources and portals?
- brings up the perennial issues:
 - what is a digital library? what are all the processes that fall under DL umbrella?



approach

- **construct** for evaluation.
 - *what was evaluated? What elements (components, parts, processes...) were involved in evaluation?*
- **context** of evaluation - selection of a goal, framework, viewpoint or level(s) of evaluation.
 - *what was the basic approach or perspective?*
- **criteria** reflecting performance as related to selected objectives.
 - *what parameters of performance were concentrated on?*
- **methodology** for doing evaluation.
 - *what measures and measuring instruments were used?*
- **findings**, except one, were **not** generalized



constructs: entities

- constructed as DL in R&D projects:
 - Perseus – classics; evaluated most
 - ADEPT – geo resources for undergrad
 - DeLIVER – sci-tech journals
 - Envision – comp. sc. literature
 - Water in the Earth System – high school
 - National Gallery of the Spoken Word - archive
 - Making of America prototype - 19th cent. journals
 - Moving Images Collection – catalog
- some are full DL, some components



constructs: entities (cont.)

- some aspect of operational DL:
 - New Zealand DL – comp. sc. tech. reports
 - ARTEMIS – science materials for school 6 to 12
 - Internet Public library – digital reference
 - UK Nat Electronic Library for Health – in a large hospital
 - Mann Library Gateway, Cornell – access interface



constructs: entities (cont.)

- multiple DL:
 - Project SOUP, Cornell – 6 digital collections in libraries & museums
 - Middlesex U – 6 general DL accessing journals & articles



constructs: entities missing

- missing evaluation of operational DLs
 - in academic, public, national libraries, museums, ...
- lot of statistics collected, but as yet not subject of evaluation
- institutional DLs are a terra incognita as to evaluation
- commercial DL products also missing from formal evaluation



constructs: processes

- variety of processes evaluated without reference to a DL:
 - various representations e.g.
 - noun-phrasing, context-based, key-phrasing
 - various tools
 - video searching, link generation, interfaces, load balancing on servers, image retrieval



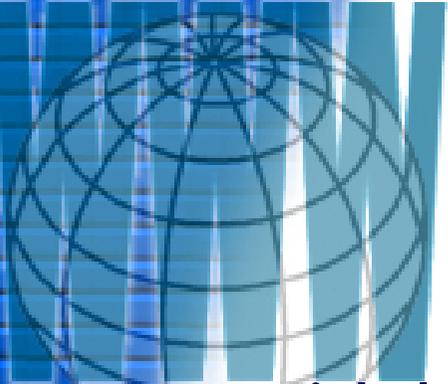
constructs: processes (cont.)

- user behavior
 - usage patterns in service logs
 - perception of quality
 - work patterns of experts
 - user preferences
 - information seeking in hypermedia DL



users: issue, borders

- when or to what extent are
user (who, why)
use (how), usage (what)
or usability studies
in DL also evaluations of DL?
- some are clearly e.g. when examining
barriers or difficulties, others are not
- is every usability study also evaluation?
- DL evaluation & studies of human
information behavior are mixed together



context of studies

- widely diverse approaches were used:
 - **Systems-centered approach:**
 - most prevalent
 - study of performance assessing effectiveness and/or efficiency
 - results may inform specific choices in design or operations
 - **Human-centered approach:**
 - also widely applied
 - study of behavior such as information seeking, browsing, searching or performance in completion of given tasks
 - implications for design, but indirectly rather than directly
 - **Usability-centered approach:**
 - assessment of different features e.g. of portals, by users.
 - a bridge between systems- and human-centered approaches.
 - mixed, or self-evident results



context of studies (cont.)

- **Ethnographic approach:** comprehensive observation of
 - life-ways, culture and customs in a digital library environment
 - impact of a digital library on a given community
 - applied successfully in a few studies, with illuminating results, particularly as to impact.
- **Anthropological approach:** comprehensive observation of
 - different stakeholders or communities and their cultures in relation to a given digital library
 - applied in one study with interesting results illuminating barriers between stakeholder communities.



context of studies (cont.)

- **Sociological approach:** assessment of
 - situated action or user communities in social setting of a DL
 - applied in one study with disappointing results
- **Economic approach:** study of
 - costs, cost benefits, economic values and impacts.
 - strangely, it was applied at the outset of digital library history (e.g. project PEAK) but now the approach is not really present at all



context of studies (cont.)

- levels of evaluation vary from
 - micro level – e.g. fast forward for video surrogates
 - macro level – e.g. impact of Perseus on the field and education in classics
- temporal aspects
 - some obsolete fast e.g. on technology
 - other longitudinal



criteria

- chosen standard(s) to judge thing by
 - there is no evaluation without criteria
 - in IR: relevance is basic criterion
 - in libraries: fairly standardized
 - in DL: no basic or standardized criteria, no agreement
 - DL metrics efforts not yet fruitful
 - thus, every evaluator choose own criteria
 - as to DL evaluation criteria
- there is a jungle out there



usability criteria

- “extent to which a user can achieve goals with effectiveness, efficiency & satisfaction in context of use” (ISO)
- widely used, but no uniform definition for DL
- general, meta criterion, covers a lot of ground
- umbrella for many specific criteria used in DL evaluations



usability criteria (cont.)

Content (of a portal or site)

- accessibility, availability
- clarity (as presented)
- complexity (organization, structure)
- informativeness
- transparency
- understanding, effort to understand
- adequacy
- coverage, overlap,
- quality, accuracy
- validity, reliability
- authority

Process (carrying out tasks as search, browse, navigate, find, evaluate or obtain a resource)

- learnability to carry out
- effort/time to carry out
- convenience, ease of use
- lostness (confusion)
- support for carrying out
- completion (achievement of task)
- interpretation difficulty
- sureness in results
- error rate



usability criteria (cont.)

Format

- attractiveness
- sustaining efforts
- consistency
- representation of labels (how well are concepts represented?)
- communicativeness of messages

Overall assessment

- satisfaction
- success
- relevance, usefulness of results
- impact, value
- quality of experience
- barriers, irritability
- preferences
- learning



systems criteria

- as DL are systems, many traditional systems criteria used
- pertain to performance of given processes/algorithms, technology, or system overall



systems criteria (cont.)

- **Process/algorithm performance**

- relevance (of obtained results)
- clustering
- similarity
- functionality
- flexibility
- comparison with human performance
- error rate
- optimization
- logical decisions
- path length
- clickthroughs
- retrieval time

Technology performance

- response time
- processing time, speed
- capacity, load

Overall system

- maintainability
- scalability
- interoperability
- sharability
- costs



other criteria

use, usage

- usage patterns
- use of materials
- usage statistics
- who uses what, when
- for what reasons/decisions

ethnographic...

in different groups:

- conceptions, misconceptions
- practices
- language, frame of reference
- communication
- learning
- priorities
- impact



methodologies

- DL are complex entities
 - many methods appropriate
 - each has strengths, weaknesses
- range of methods used is wide
 - there is no “best” method
 - but, no agreement or standardization on any methods
- makes generalizations difficult, even impossible



methodologies (cont.)

- surveys
- interviews
- observations
- think aloud
- focus groups
- task performance
- log analysis
- usage analysis
- record analysis
- experiments
- economic analysis
- case study
- ethnographic analysis



results

- not synthesized here
- hard to synthesize anyhow
- generalizations are hard to come by
- except one!



users and digital libraries

- a number of studies reported various versions of the same result:

users have many difficulties with DLs

- usually do not fully understand them
- they hold different conception of a DL from operators or designers
- they lack familiarity with the range of capabilities, content and interactions
- they often engage in blind alley interactions



analogy

- perceptions of users and perceptions of designers and operators of a DL are generally not very close
- users are from Venus and DLs are from Mars
- leads to the **versus** hypothesis



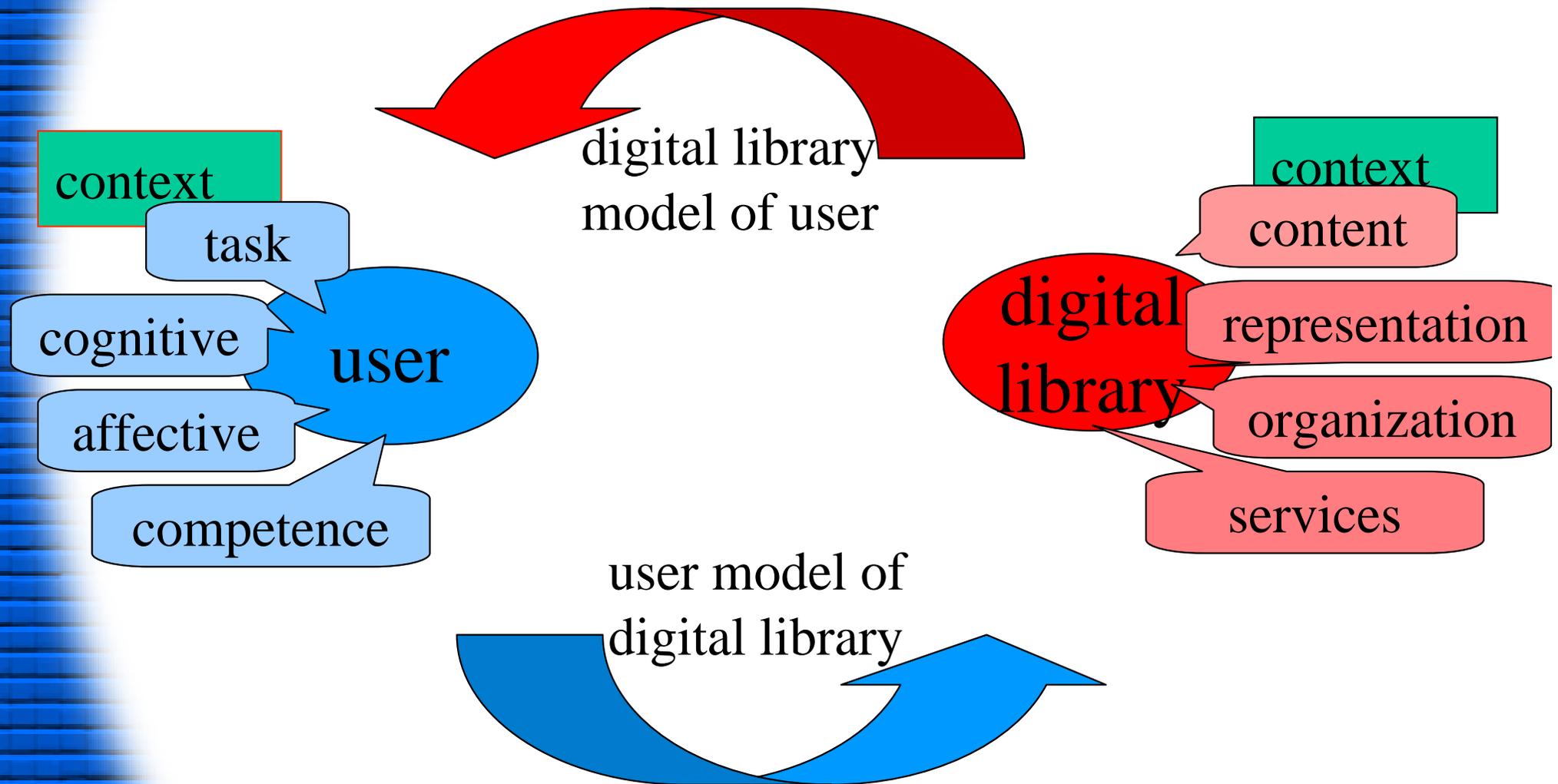
is it:

user AND digital library
or
user VERSUS digital library

- why VERSUS?
 - users and digital libraries see each other differently



user **AND** digital library model





how close are they?

user **VERSUS** digital library model



user model of digital library

what user assumes about digital library:
how it works?
what to expect?

digital library model of user

what digital library assumes about user:
- behavior?
- needs?



the **versus** hypothesis

in use, more often than not, digital library users and digital libraries are in an adversarial position

- hypothesis does not apportion blame
 - does not say that DL are poorly designed
 - or that users are poorly prepared
- adversarial relation may be a natural order of things



evaluation of digital libraries

- impossible? not really
- hard? very
- could not generalize yet
- no theories
- no general models embraced yet, although quite a few proposed
- in comparison to total works on DL, only a fraction devoted to evaluation



why? – some speculations

- **Complexity:** DLs are highly complex
 - more than technological systems alone
 - evaluation of complex systems is very hard
 - just learning how to do this job
 - experimenting with doing it in many different ways
- **Premature:** it may be too early in the evolution of DL for evaluation on a more organized scale



why? (cont.)

- **Interest:** There is no interest in evaluation
 - R&D interested in doing, building, implementing, breaking new paths, operating ...
 - evaluation of little or no interest, plus there is no time to do it, no payoff
- **Funding:** inadequate or no funds for evaluation
 - evaluation time consuming, expensive requires commitment
 - grants have minimal or no funds for evaluation
 - granting agencies not allocating programs for evaluation
 - no funds = no evaluation.



why? (cont.)

- **Culture:** evaluation not a part of research and operations of DL
 - below the cultural radar; a stepchild
 - communities with very different cultures involved
 - language, frames of reference, priorities, understandings differ
 - communication is hard, at times impossible
 - evaluation means very different things to different constituencies



why – the end

- **Cynical:** who wants to know or demonstrate actual performance?
 - emperor clothes around?
 - evaluation may be subconsciously or consciously suppressed
 - dangerous?



ultimate evaluation

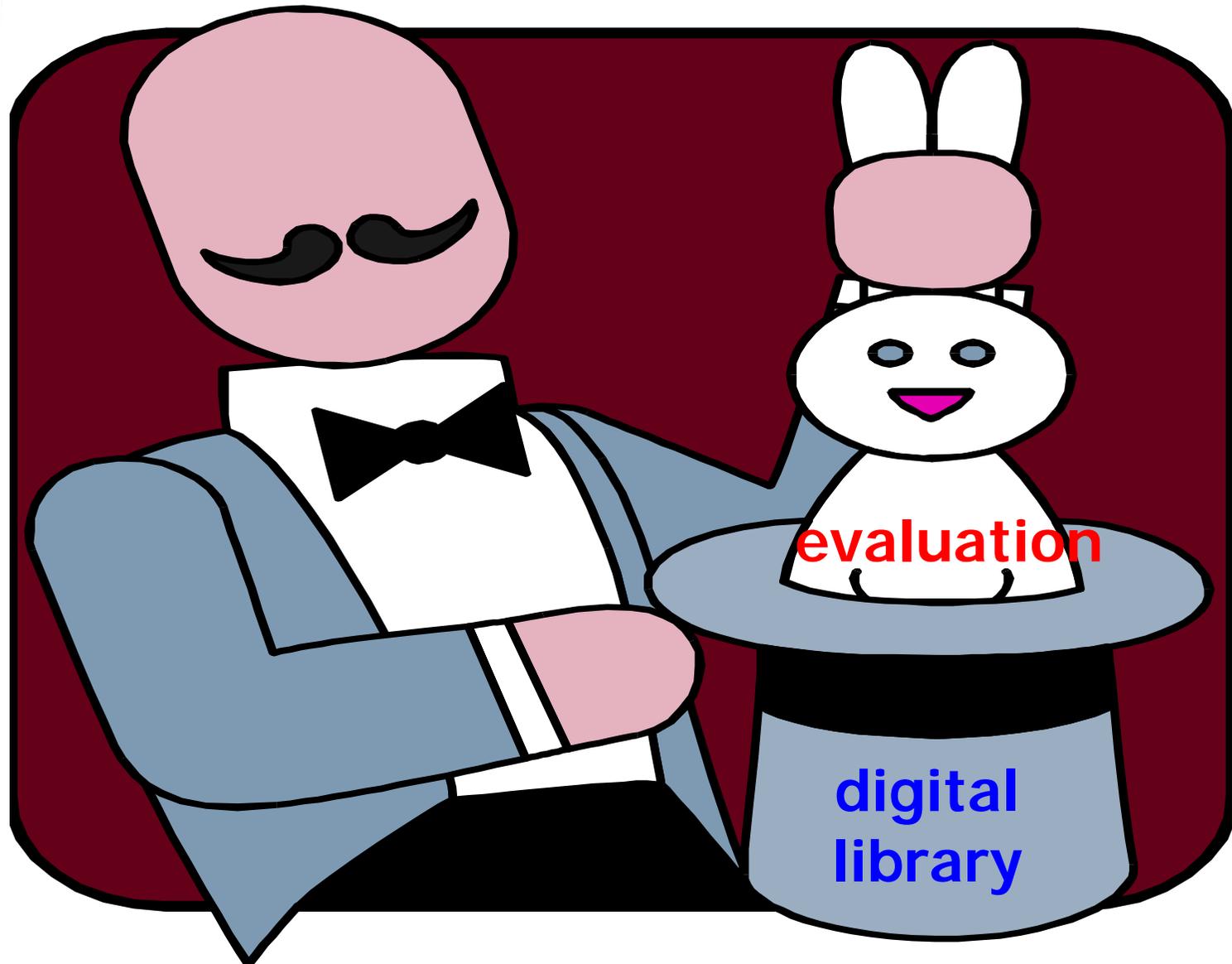
- The ultimate evaluation of digital libraries:
 - assessing transformation in their context, environment
 - determining possible enhancing changes in institutions, learning, scholarly publishing, disciplines, small worlds ...
 - and ultimately in society due to digital libraries.



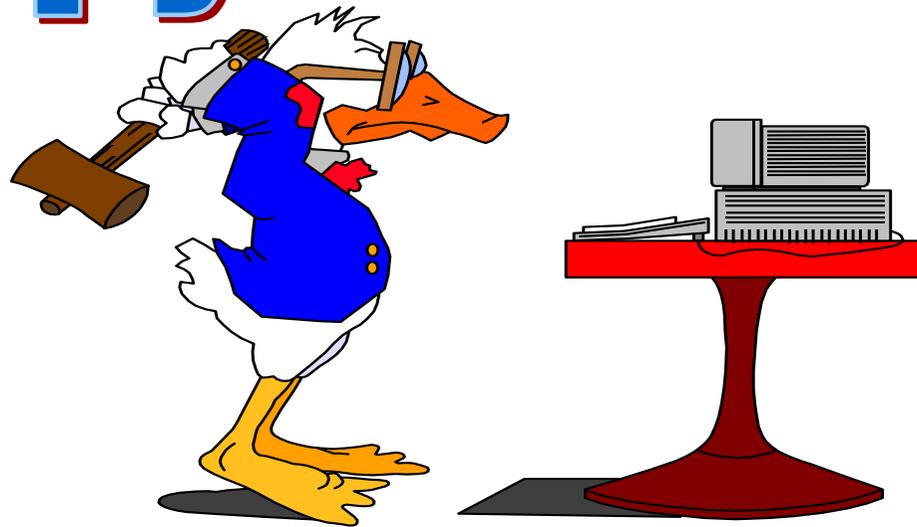
conclusions

- evaluation of digital libraries still in formative years
- not funded much, if at all
- but necessary for understanding how to
 - build better digital libraries & services &
 - enhance their role

How to do it?



Happy evaluation!





hvala

grazie
thank you

ευχαριστίες

danke

tak

takk

tack

köszönöm



sources

- the paper and PowerPoint presentation at:

<http://www.scils.rutgers.edu/~tefko/articles>

- annotated bibliography at:

<http://www.scils.rutgers.edu/~miceval>