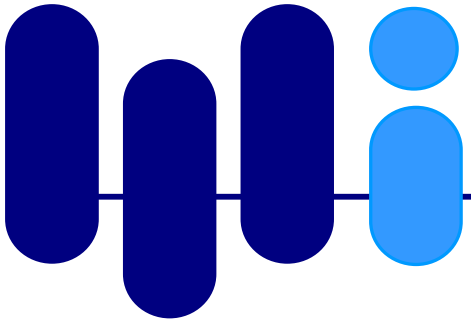
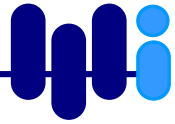


# Information Retrieval is for Everybody

FGIR Workshop
Hildesheim University, October 2006

*Christian Wolff*
*Regensburg University, Media Computing*



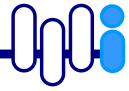
# Overview

- Motivation
- Development phases in Information Retrieval
- Users and information literacy
- A storytelling example: A day in the life of the common IRS user
- Some observations and implications for IR research
- Conclusion

# Motivation

- IRS usage commonplace in everyday life
- multimedia and mobile computing call for new IRS application (e.g. GIS, digital photography, ubiquitous network access)
- Web 2.0 challenges assumptions on user participation (tagging, user involvement)
- exploitation of
  - contextual information
  - new interaction techniques

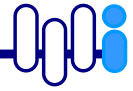
# Phases in IRS Development



1. IRS as instruments of information professionals in mediated communication contexts
2. IRS used by knowledge workers ("end users")
3. web-based search: IRS used for (almost) anything (desktop setting)
4. "the digitization of the world picture" (Cerruzzi 2003): information systems and IRS are becoming ubiquitous, mobile, context-aware …

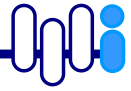
(phases overlap heavily!)

# Users and information literacy



- ARD online survey 2005:
  60% of the German population regularly use the internet, more than 90% in the group of the 14-19years old
- almost 100% of pupils and students have private internet access (small studies with ~300 participants in 2005/2006)
- majority of children is familiar with basic internet and search engine concepts
- evident deficits in information literacy
  - linguistic phenomena
  - search strategies
  - query languages and operators
  - information quality judgements
  - knowledge about available resources

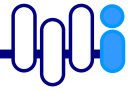
# A storytelling example: IRS in everyday life



- In the morning: Collecting music, loading the iPod
  - different genre classifications
  - retrieval by example not viable for many tasks (humming, singing, whistling)
  - problems of describing and classifying music
  - limited portability of licenses
  - no similarity search (see MusicFinder)
  - problem of information management

# Example: ID3-Tagging a Music Library

# A storytelling example: IRS in everyday life

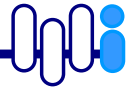


- In the morning
- At work:
  - searching for literature
    - heterogeneous tools and information
    - different retrieval models and data source characteristics
    - complex criteria for availability of sources
  - personal information management
    - limitations of traditional file system structures (monohierarchical)
    - traditional fulltext retrieval becoming available via desktop search engines (classic VSM paradigm)
    - advanced techniques (aspect/context/situation awareness, embedded retrieval) missing

# Matrix of typically available literature access systems

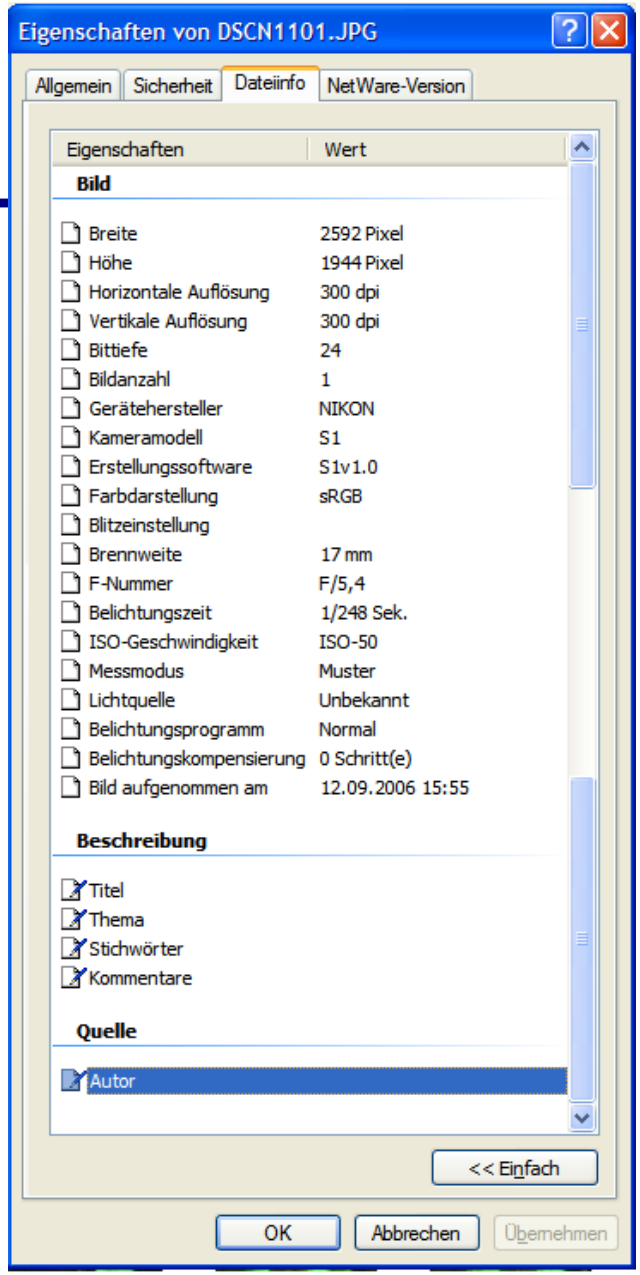
| | **Finding** | **Accessing / Ordering** |
|---|---|---|
| books (monographs, edited books) | local OPAC, regional catalogs, book sellers („Amazon“) | local OPAC, regional catalogs, book sellers („Amazon“) |
| papers (journals, conferences) | Integrated DBs (Scopus, Web of science), scince DBs (INSPEC, | electronic journal library, publishers‘ dls, integrated portals (vascoda, io-port, ...) |
| grey literature | Citeseer, Google Scholar, integrated portals | directly online |
| primary research data, software, ... | Google, specialised dbs | online? (registration?) |

# A storytelling example: IRS in everyday life

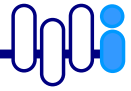


- In the morning
- At work
- In the afternoon
  - online shopping
    - great number of platforms and meta platforms
  - „a colleague‘s visit“: taking (digital) pictures
    - manyfold media
    - no content metadata on media production time

# „a colleague‘s visit“: no descriptive metadata

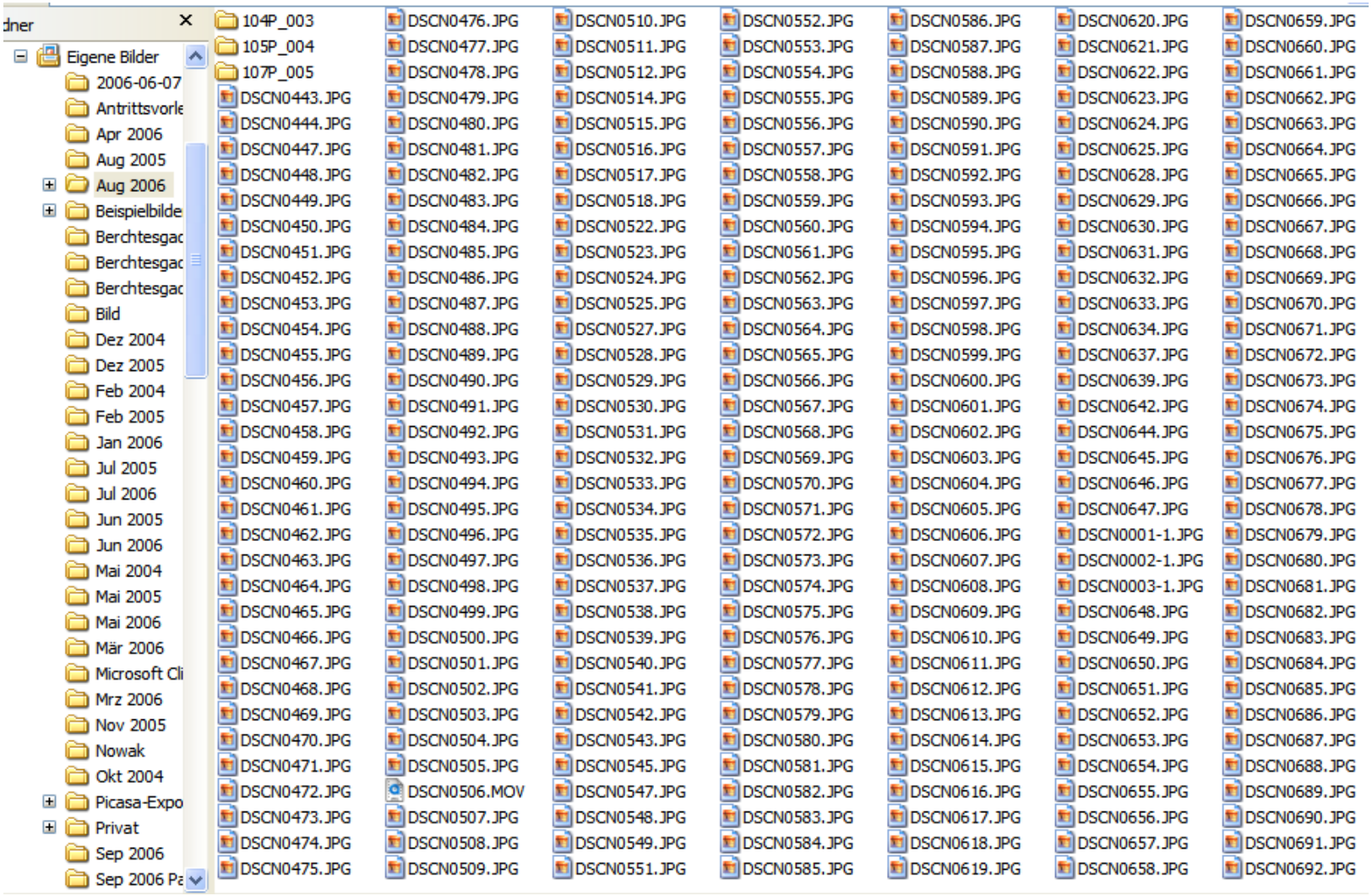


# A storytelling example: IRS in everyday life

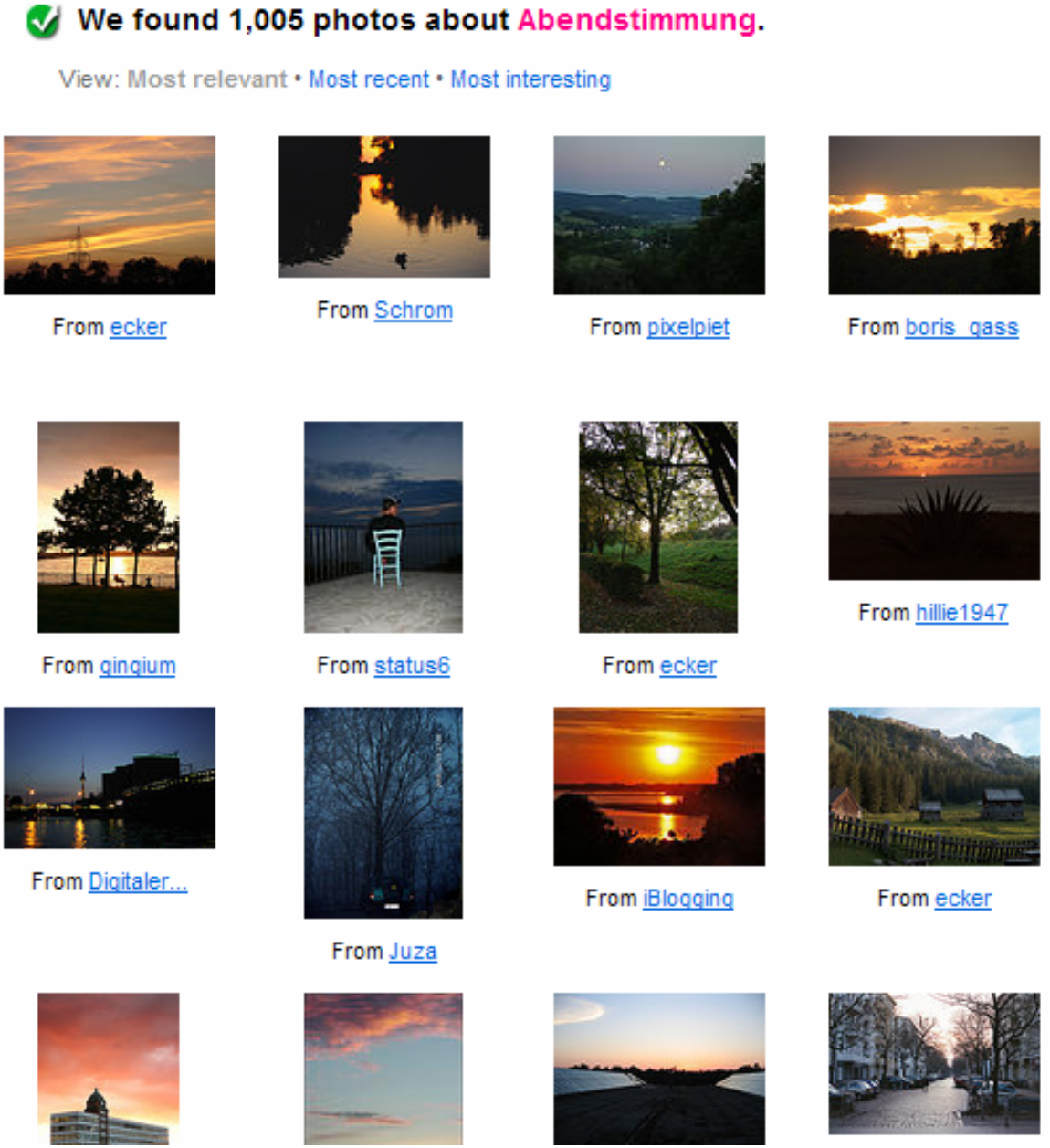


- In the morning
- At work
- In the afternoon
- In the evening
  - some time for tagging and picture sorting, querying images in Flickr
  - (media convergence: interactive TV and web-based IS)

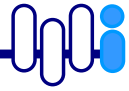
# What does a typical folder with holiday pictures look like?



# „Abendstimmung“



# A storytelling example: IRS in everyday life



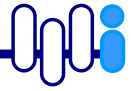
- In the morning
- At work
- In the afternoon
- In the evening
- At night
  - desktop index is being updated
  - image analysis compares pictures and suggests tags
  - user interaction data are evaluated for further search
  - (the user sleeps)

# What is new? - IR and everyday life as objects of science – is this

- ***legitimate***?
  - „Information Retrieval (IR) deals with the representation, storage, organization of, and access to information items. The representation and organization of the information items should provide the user with easy access to the information in which he is interested." (Baeza-Yates & Ribeiro-Neto 1999:1)
- ***different*** – or just the same research paradigm?
  - different
    - users (age, training, interests),
    - situations and contexts,
    - types of information / data (quality) …

  Shneiderman's „the new computing is about what people can do" (Sheiderman 2002:2) calls for a scientific approach to computing in everyday life, including IR research

# Some theses



- **everybody** is an IRS user
- the **amount of data** (files, information knowledge) is growing rapidly (e.g. > 500.000 items in my Google desktop index)
- **all media** are subject to IR processes - in the media production chain not just ***search*** is relevant, but also the ***descriptive*** steps (indexing, tagging)
- **economic**, **organizational** as well as **temporal** criteria influence effectiveness measures (as compared to traditional effectiveness)
- the potential of **social software** is still quite unclear
- **text** and **concept related query paradigm** will prevail for a while and for most media
- **media convergence** will gain importance

# Some questions ...

- new evaluation criteria needed?
  - user satisfaction
  - task completion
- how do we
  - describe
  - analyse

  situation, context, interaction history etc.?
- how do we fuse different data sources like
  - declarative knowledge (on the user)
  - sensor data
  - ...

# Conclusion

- demand for
  - user-related research
  - information literacy in everyday life
    - taking IL beyond the library user / student paradigm
    - the digital divide as a problem
      - not of access to technology
      - but as of a lack of training
  - models
    - information in everyday life
    - of future I(R)S
    - modeling context/situation/experience

# Examples: „the knowing camera“

- goal: multi-source information collection in digital photography
- technology:
  - speech interaction with digital camera for metadata generation
  - sensors for place (GPS), angle/viewpoint
  - (temperature, humidity, biodata ...)
  - (machine learning, pattern recognition)
- applications
  - enter-/infotainment

# Examples: exploring tagging strategies

- research questions
  - tagging strategies in
    - professional sources
    - social software platforms
- method
  - analyzing samples form Flickr / Citeulike
  - building a tag classification
  - content overlap (Citeulike)
- meta-tagging tags based on that classification
- comparative study: search with / without tagged documents

# Some Literature


- Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval. Harlow et al. / New York: Addison-Wesley / ACM Press.*
- Ceruzzi, P. E. (2003). *A History of Modern Computing (2nd ed.). Cambridge, MA / London: The MIT Press.*
- Shneiderman, B. (2002). *Leonardo's Laptop: Human Needs and the New Computing Technologies. Cambridge, MA / London: The MIT Press.*
- van Eimeren, B., & Frees, B. (2005). ARD/ZDF-Online-Studie 2005. Nach dem Boom: Größter Zuwachs in internetfernen Gruppen. *Media Perspektiven(8/2005), 362-379.*