Advanced Multimedia

Meta Data & Tags
Tamara Berg

Some slides adapted from: JISC Digital Media
Reminders

• HW4 due tonight
• Project Proposal Presentations on Thurs
  – If you haven’t discussed your project idea with me please come to office hours today
  – Please prepare a 5 minute presentation outlining:
    • Project idea
    • Challenges to be overcome
    • Timeline for accomplishing challenges
What is metadata?

Metadata is often defined as ‘data about data’. In the case of digital assets, metadata can be textual information that describes something about the creation, content, or context of an image (e.g. date made, subject matter, location of digital file).

Kinds of metadata:
- Controlled terminology, carefully constructed or chosen from formal lists and entered into pre-established categories.
- Or it might be a simply free text description or set of keywords used to annotate or ‘tag’ a digital resource.
- It might describe something objective and straightforward, such as the file size of the digital file; or something much more complex, such as the subject matter of the resource or legal rights associated with its use.
- Metadata is often held within databases, but it can take other forms - it can just as easily be found embedded within the digital file itself.
Selective

- Metadata invariably offers a selective or simplified description of a resource.
- The Oxford English Dictionary defines metadata as “data that operates at a higher level of abstraction”. If “a picture paints a thousand words” (or more!) it is clear that our text-based descriptions will only ever partially capture the information or meanings held within a digital resource - let alone all the other information that might be associated with it (e.g. the history of its creation, its relationship to other resources, or possible uses to which it might be put).
- The challenge for those applying metadata to a digital collection is to work out which information is going to be the most important and useful to record.
Structured metadata

• Metadata can be structured in some way.
• Rather than randomly associating terms with the digital file, it is common to use a set of generic categories and then assign specific terms within those categories.
### Structured metadata

<table>
<thead>
<tr>
<th>Metadata schema categories</th>
<th>Metadata vocabulary terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creator</td>
<td>Leonardo da Vinci</td>
</tr>
<tr>
<td>Title</td>
<td>Mona Lisa</td>
</tr>
<tr>
<td>Subject</td>
<td>woman, portrait, Renaissance...</td>
</tr>
<tr>
<td>...etc</td>
<td></td>
</tr>
</tbody>
</table>
Structured Metadata

This approach has several advantages:

* Easy to create the metadata, since the categories tell the cataloguer which information needs to be collected and recorded.
* Easy to understand the metadata, making it clear to a user, for example, that it is Leonardo who has created this image rather than Mona Lisa!
* Easy to retrieve the image in a search, since the search query can be much more specific, targeting relevant categories rather than searching across all of the metadata.
* Easy to share the image and its metadata with other image collections – as long as common categories and terminologies have been used.

Another phrase you will often find used in these papers (and elsewhere) is “controlled vocabularies.” This is used where the specific terminology used within a metadata category has been drawn from a pre-defined list (e.g. thesauri) or has been constructed according to a standard set of rules (e.g. “enter the creator name in this form: ‘Surname, Forenames’”).
Structured Metadata

This approach has several advantages:
* Easy to create the metadata, since the categories tell the cataloguer which information needs to be collected and recorded
* Easy to understand the metadata, making it clear to a user, for example, that it is Leonardo who has created this image rather than Mona Lisa!
* Easy to retrieve the image in a search, since the search query can be much more specific, targeting relevant categories rather than searching across all of the metadata
* Easy to share the image and its metadata with other image collections - as long as common categories and terminologies have been used
Structured Metadata

This approach has several advantages:
* Easy to create the metadata, since the categories tell the cataloguer which information needs to be collected and recorded
* Easy to understand the metadata, making it clear to a user, for example, that it is Leonardo who has created this image rather than Mona Lisa!
* Easy to retrieve the image in a search, since the search query can be much more specific, targeting relevant categories rather than searching across all of the metadata
* Easy to share the image and its metadata with other image collections - as long as common categories and terminologies have been used

Another phrase you will often see is “controlled vocabularies”. This is used where the specific terminology used within a metadata category has been drawn from a pre-defined list (e.g. thesauri) or has been constructed according to a standard set of rules (e.g. “enter the creator name in this form: ‘Surname, Forenames’”).
Purposes

- Metadata can serve different purposes:
  - Help us to find the resource (termed 'resource discovery' metadata).
  - Tell us what it is (descriptive metadata).
  - Tell us where the resource has come from, who owns it and how it can be used (provenance and rights metadata).
  - Describe how the digital resource was created (technical metadata).
  - How it is managed (administrative metadata).
  - How it can be kept into the future (preservation metadata).
  - Help us to relate this digital resource with other resources (structural metadata).
Purposes

• Metadata can serve different purposes:
  – Help us to find the resource (often termed ‘resource discovery’ metadata).
  – Tell us what it is (descriptive metadata).
  – Tell us where the resource has come from, who owns it and how it can be used (provenance and rights metadata).
  – Describe how the digital resource was created (technical metadata).
  – How it is managed (administrative metadata).
  – How it can be kept into the future (preservation metadata).
  – Help us to relate this digital resource with other resources (structural metadata).
Communities

• Metadata does not exist in a vacuum, it serves particular purposes and particular groups of users. There will often be different kinds of users with different needs. In developing a metadata framework for your collection it is important that you identify all of these users and needs. It’s best to ask your users what information they need rather than make assumptions.

• Digital collections are often based within particular ‘communities’, for example: libraries, archives, museums, educators. Many of the formal metadata standards currently in use have been developed within such communities.
  – Means that these standards are generally good at supporting the needs of that community.
  – However, they can also incorporate old-fashioned or ‘legacy’ approaches that may have worked well in a non-digital environment, but are not as practical or useful in the digital world.
Where does metadata come from?

• Metadata relating to a digital resource can come from one of two sources: (a) it can be automatically derived from the digital resource itself, or (b) can be created and associated with a resource by human beings.

(a) Intrinsic or implicit metadata (file formats, resolution, bit-depth, or colourspace). File formats typically encode this sort of information within the header of the digital file. If an image has been created by a digital camera, it is likely that the camera has also written a certain amount of information about the digital capture into the file header (EXIF standard), such as the camera make and model, its settings, and the date the photograph was taken.

Most implicit metadata is technical in nature and is generally of more use to those administering the collection rather than those using it.

In developing a digital collection it may be useful to extract some implicit metadata and hold it separately within a database for the purposes of retrieval, quality control, or digital preservation.
Where does metadata come from?

(b). extrinsic or explicit metadata. Because this is created by humans, it is the most difficult and expensive metadata to create. But it is also usually the most important - especially to the end user.

Although explicit metadata must be created by humans, it need not all be created by those building and cataloguing a digital collection. It is very likely that there is some pre-existing legacy metadata that can be exploited (even if it is just a scrawled inscription on the back of a photograph, film can or audio cassette). Or it might be possible to get your collection users to add to the metadata in some way (via tags or annotations).
Where is metadata kept

• Metadata for digital collections can be held in several different places: (a) within the digital file; (b) within a database; (c) in a separate XML-encoded file; or (d) all of the above at once.

• As the earlier discussion of implicit metadata has indicated, there is already a certain amount of metadata held within a digital file. Some of this might be extracted for use outside of the digital asset. In addition to extracting metadata, it is possible to embed some metadata within the digital asset.

• Most people developing digital collections will make use of a database to hold their metadata.
XML

- Increasingly, XML is being used as a way of encoding metadata. XML is related to HTML (the original coding used on the World Wide Web). While the HTML tags are primarily focused on presentation (e.g. `<b>Bold</b>`), XML tags are used to indicate meaning (e.g. `<organisationName>JISC Digital Media</organisationName>`). This approach lends itself well to expressing metadata, enabling metadata schema categories to be turned into tags and wrapped around specific terms.

```
<image record>
  <original work>
    <format>painting</format>
    <creator>Leonardo da Vinci</creator>
  </original work>
  <reproduction>
    <format>photographic transparency</format>
    <creator>Jane Smith</creator>
  </reproduction>
  <reproduction>
    <format>JPEG image</format>
    <creator>John Brown</creator>
  </reproduction>
</image record>
```
HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, ToRead

Cameron Marlow, Mor Naaman, Danah Boyd, Marc Davis
Web based tagging systems

Del.icio.us, Technorati, Flickr

Allow participants to annotate a particular resource (web page, blog post, image, physical location, ...)

• Unstructured
• No controlled vocabulary
• User annotated
Example tagging systems

Del.icio.us (http://del.icio.us): a “social bookmarking site,” allowing users to save and tag web pages and resources.

Yahoo! MyWeb2.0 (http://myweb.yahoo.com): similar to Del.icio.us, but including a social network of contacts.

CiteULike (http://www.citeulike.org/): a site allowing users to tag citations and references, e.g. academic papers or books.

Flickr (http://www.flickr.com): a photo sharing system allowing users to store and tag their personal photos, as well as maintain a network of contacts and tag others photos.

YouTube (http://www.youtube.com): a video sharing system allowing users to upload video content and describe it with tags.

ESP Game (http://www.espgame.org/) [23]: an internet game of tagging where users are randomly paired with each other, and try to guess tags the other would use when presented with a random photo.

Last.fm (http://www.last.fm): a music information database allowing members to tag artists, albums, and songs.

Yahoo! Podcasts (http://podcasts.yahoo.com/): a site that indexes podcasts (regularly updated audio content), and allows users to tag them.

Odeo (http://www.odeo.com/): another podcast information system supporting tagging and search.

Technorati (http://www.technorati.com/): a weblog aggregator and search tool allowing blog authors to tag their posts.

LiveJournal (http://www.livejournal.com/): a weblog and community website allowing users to tag their personal profile, along with individual blog posts.

Upcoming (http://upcoming.org/): a collaborative events database where users can enter future events (e.g., concerts, exhibits, plays, etc.) and tag them.
Tagging as Social bookmarking

• Tagging a resource is similar to categorizing personal bookmarks.
  – Tags allow users to store and collect resources and retrieve them using the tags applied
  – Similar keyword-based systems have existed in web browsers, photo repository apps, and other collection management systems
  – What’s new is the element of social interaction – connecting bookmarking activities to a network of shared tags, resources, and users.
Social Tagging systems

• Allow users to share their tags for resources
• Tags serve as links to other resources tagged the same way
• Rely on emergent social structures and behaviors (since there is no pre-defined vocab – folksonomy).
• Shared pool of tags
Social Tagging systems

- Allow users to share their tags for resources
- Tags serve as links to other resources tagged the same way
- Rely on emergent social structures and behaviors (since there is no pre-defined vocab – folksonomy).
- Shared pool of tags
  - Enhances metadata for all users
  - Distributes workload for metadata creation among many contributors.
  - May overcome the problems of having no pre-defined vocab through lots of data!
Model of tagging systems

Resources (web pages, images etc) - may be linked to each other through hyperlinks etc.

Users – May be linked to each other through social ties and social networks.

Tags – Edges connecting users with resources. Tradeoffs between crafted ontologies and folksonomies.
Challenges of tagging systems

• Semantic difficulties
  – Polysemy – single word may have multiple meanings
  – Synonymy – different words may have the same meaning
  – Different expertise and purposes of users may result in varying levels of abstraction.
    • Basic level – cat
    • Superordinate level – animal
    • Subordinate levels – persian cat, felis silvestris catus longhair persian.
Labeled photos

Sense ambiguity (tiger):

Disconnect b/w tagger and searcher purposes (tiger):
System Design Considerations
Tagging Rights

• System’s restriction on group tagging
  – Self-tagging - only allow users to tag resources they created (technorati).
  – Free-for-all tagging – any user can tag any resource (Yahoo! Podcasts).
  – Varying levels of compromise
    • Choose the resources users are to tag (images in the ESP game)
    • Specify different levels of permission to tag (friends, family, and contacts in Flickr).
    • System can determine who may remove a tag, no one (Yahoo! Podcasts), anyone (Odeo), tag creator (Last.fm) or resource owner (Flickr).
Tagging support

- **Blind tagging** – a tagging user cannot view tags assigned to the resource by other users while tagging (Del.icio.us)
- **Viewable tagging** – user can see the tags already associated with a resource (Yahoo! Podcasts)
- **Suggestive tagging** – system suggests possible tags to the user (Yahoo! MyWeb2.0)
  - Suggested tags may be based on existing tags by same user, tags assigned to same resource by others, from related tags (machine-suggested tag synonyms).
Aggregation

• Aggregation of tags around a given resource
  
  Bag-model – allow for multiplicity of tags for same resource which may result in duplicate tags from diff users (Del.icio.us)
  
  Set-model – ask the group to collectively tag a resource, denying any repetition (Flickr)
Type of Object

• Type of resource being tagged:
  – Web pages (Del.icio.us, Yahoo! MyWeb2.0)
  – Bibliographic material (CiteULike)
  – Blog posts (Technorati, LiveJournal)
  – Video (YouTube)
  – Audio objects (Last.fm)
  – Podcasts (Yahoo! Podcasts, Odeo)
  – Locations or events (Upcoming)
  – Images (Flickr)
Source of Material

Supplied by participants (YouTube, Flickr, Technorati, Upcoming)

Supplied by the system (ESP Game, Last.fm, Yahoo! Podcasts)

Open to tagging any web resource (Del.icio.us, Yahoo! MyWeb2.0)

Some systems restrict the source through architecture (Flickr)

Others restrict the source solely through social norms (CiteULike).
Model of tagging systems

Resources (web pages, images etc) - may be linked to each other through hyperlinks etc.

Users – May be linked to each other through social ties and social networks.

Tags – Edges connecting users with resources. Tradeoffs between crafted ontologies and folksonomies.
Resource Connectivity

• Resources can be linked to each other independent of the user tags
  – Linked – web pages connected by links
  – Grouped – Flickr photos can be assigned to groups, events in Upcoming have connections based on time, city and venue.
  – None – resources don’t have to be otherwise linked together.
Social Connectivity

Some systems allow the users to be linked together
- Linked
- Grouped
- None

Links can be typed (Flickrs contacts/friends model), directed (where connection between users is not necessarily symmetric).
Why do users tag?
Future retrieval

• To mark items for personal retrieval of the individual resource or collection of clustered resources (tagging a group of papers on Del.icio.us in preparation for writing a book, tagging songs on Last.FM to create a playlist, tagging Flickr photos ‘home’ to be able to find all photos taken at home later).

• These tags may also be used to incite an activity or serve as reminders (“to read” tag).

• These descriptive tags are exceptionally helpful in providing metadata about objects that have no other tags associated.
Contribution & Sharing

• To add to conceptual clusters for the value of either known or unknown audiences
  – Tag vacation websites for a partner
  – Contribute concert photos and identifying tags to Flickr for anyone who attended the show
Attract Attention

• To get people to look at one’s own resource
  – To be returned at the top of results by search queries
  – When “tag clouds” or other such lists that reflect popularity of tags are visible in the system, users may be incentivized to contribute tags that might affect that global view (ie use common tags)
    • Why might this be a bad thing?
Play and Competition

• To produce tags based on an internal or external set of rules.
  – System might devise the rules (ESP game’s incentive to tag what others might also tag)
  – Groups might develop their own rules to engage in the system such as when groups seek out all items with a particular feature and tag their existence or get the owner to add the resource to the group

http://www.gwap.com/gwap/gamesPreview/espgame/
Self Presentation

• To write a user’s own identity into the system as a way of leaving their mark on a particular resource
  – E.g. “seen live” tag in Last.FM marks an individual’s identity or personal relation to the resource.
Opinion Expression

• To convey value judgments that they wish to share with others.
  – E.g. “elitist” tag in Yahoo!’s Podcast system is utilized by some users to convey an opinion
Tag Types

• Properties of objects (could arise from organizational motivations, social ones or both)
  – Source
  – Attributes
  – Category membership
  – Qualitative properties

• Self-reference
  – Mystuff, mywork reflects an intent to communicate ownership to an outside audience or alternatively to be used for personal organization

• Task-Organization
  – Toread, jobsearch suggests an intent for personal organization.
Case Study: Flickr
All about Flickr

8.5 million registered users
Up to 12,000 photos served per second
Up to 3 million photos uploaded per day
Over 3 billion photos total
All about Flickr

• Photo sharing site
• Considers tags as a core element to sharing, retrieval, navigation, and discovery of user-contributed images.
• Allows users to upload their personal photos to be stored online, but makes them publicaly viewable and easily discoverable by default.
• This design decision, along with emphasis on tagging has allowed the site to expand quite rapidly over its short lifespan.
Title: LAKE COMO

Tags: ossuccio, lakecomo, como, lake, water, shadow, door, light, mountains, snow, sun, italy, holiday, europe

Comments:
EMCphotos says:
Excellent shot.
A 2 Z Member Choice!

ickleFlower says:
Totally Inspired
What a Great Take

predestinariane says:
Remarkable.....

Description: Ossuccio - Santa Maria Maddalena

Photo by newpn2000
Tags on Flickr

• Tags are user-contributed
• Tagging rights are restricted to self-tagging or permission-based
• Tags are aggregated in sets (as opposed to bags).
• Interface mostly affords blind-tagging.
Tags on Flickr

• Flickr users are likely to tag for their own retrieval
• But the system also encourages gaming and exploration of tag use
  – Users are primarily motivated by social incentives, including the opportunities to share.
Tag Usage on Flickr

- Tags are not mandatory, but an optional feature in a larger resource organization task.
- Flickr interface prompts users for metadata about each resource identified: title, caption, list of tags.
- System allows users to tag their friends’ photos, but this feature is not largely used; overwhelming majority of tags are applied by the photo owners.
Tag Usage Patterns

Probability that a random user has a set of distinct tags of size $k$.

e.g. probability that a Flickr user has more than 750 distinct tags is about 0.1%
Tag Usage Patterns

Probability that a random user has a set of distinct tags of size k.

e.g. probability that a Flickr user has more than 750 distinct tags is about 0.1%

Most users have very few distinct tags, while a small group has extremely large sets of tags.
Flickr usage correlation

Table 2. Flickr usage correlation

<table>
<thead>
<tr>
<th></th>
<th>Tags</th>
<th>Photos</th>
<th>Contacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tags</td>
<td>1</td>
<td>.518</td>
<td>.386</td>
</tr>
<tr>
<td>Photos</td>
<td>.518</td>
<td>1</td>
<td>.192</td>
</tr>
<tr>
<td>Contacts</td>
<td>.386</td>
<td>.192</td>
<td>1</td>
</tr>
</tbody>
</table>

* N = 25,000
** p < 0.001 for all values.

Correlation between number of tags, photos and contacts across 25,000 users.
Flickr usage correlation

Table 2. Flickr usage correlation

<table>
<thead>
<tr>
<th></th>
<th>Tags</th>
<th>Photos</th>
<th>Contacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tags</td>
<td>1</td>
<td>.518</td>
<td>.386</td>
</tr>
<tr>
<td>Photos</td>
<td>.518</td>
<td>1</td>
<td>.192</td>
</tr>
<tr>
<td>Contacts</td>
<td>.386</td>
<td>.192</td>
<td>1</td>
</tr>
</tbody>
</table>

* N = 25,000
** p < 0.001 for all values.

Correlation between number of tags, photos and contacts across 25,000 users.

Strong linear relationship between these variables.
Flickr usage correlation

Table 2. Flickr usage correlation

<table>
<thead>
<tr>
<th></th>
<th>Tags</th>
<th>Photos</th>
<th>Contacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tags</td>
<td>1</td>
<td>.518</td>
<td>.386</td>
</tr>
<tr>
<td>Photos</td>
<td>.518</td>
<td>1</td>
<td>.192</td>
</tr>
<tr>
<td>Contacts</td>
<td>.386</td>
<td>.192</td>
<td>1</td>
</tr>
</tbody>
</table>

* N = 25,000
** p < 0.001 for all values.

Correlation between number of tags, photos and contacts across 25,000 users.

Strong linear relationship between these variables.

Strongest relationship is between photos and distinct tags.
Flickr usage correlation

Table 2. Flickr usage correlation

<table>
<thead>
<tr>
<th></th>
<th>Tags</th>
<th>Photos</th>
<th>Contacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tags</td>
<td>1</td>
<td>.518</td>
<td>.386</td>
</tr>
<tr>
<td>Photos</td>
<td>.518</td>
<td>1</td>
<td>.192</td>
</tr>
<tr>
<td>Contacts</td>
<td>.386</td>
<td>.192</td>
<td>1</td>
</tr>
</tbody>
</table>

* N = 25,000
** p < 0.001 for all values.

Correlation between number of tags, photos and contacts across 25,000 users.

Strong linear relationship between these variables.

Strongest relationship is between photos and distinct tags

Weaker one between contacts and tags – suggests tagging is related to social activity to some degree.
Use of tags over time

How does the frequency of tags change as user becomes acclimated to the system?

Do her tags become a cohesive taxonomy over time, or do they continue to grow linearly as new resources are added.
How does the frequency of tags change as user becomes acclimated to the system?

Do her tags become a cohesive taxonomy over time, or do they continue to grow linearly as new resources are added.
Use of tags over time

How does the frequency of tags change as user becomes acclimated to the system?

Do her tags become a cohesive taxonomy over time, or do they continue to grow linearly as new resources are added.

Growth of distinct tags for 10 randomly selected users – selected as frequent uploaders (>100 photos), and frequent taggers (>100 tags).

Behavior varies by user:
Some cases new tags are added consistently as photos are uploaded (A).
Sometimes only a few tags are used initially with a sudden growth spurt later on (B).
For many users, distinct tag growth declines over time.
Possible explanations

• Is the linear growth related to the type of media being tagged, namely photos that are taken of constantly evolving subject matter?
• Or does it evolve from a motivation to continually attract new individuals to the users’ photos?
• A sudden increase could reflect a change in personal motivations (ie a need to start organizing photos once the collections grows above a certain size), or a social one (e.g. a sudden realization that tags can bring new people to see one’s photos).
Popular Tags

africa amsterdam animals april architecture art asia australia baby band barcelona beach berlin bird birthday black blackandwhite blue boston bw california cameraphone camping canada canon car cat chicago china christmas church city clouds color concert cute dance day de dog england europe fall family festival film florida flower flowers food france friends fun garden geotagged germany girl girls graffiti green halloween hawaii hiking holiday home honeymoon house india ireland island italy japan july kids la lake landscape light live london macro me mexico mountain mountains museum music nature new newyork newyorkcity newzealand night nikon nyc ocean paris park party people photo photos portrait red river rock rome san sanfrancisco scotland sea seattle show sky snow spain spring street summer sun sunset sydney taiwan texas thailand tokyo toronto travel tree trees trip uk urban usa vacation vancouver washington water wedding white winter yellow york zoo
Rare tags

Returned by tag search for “monkeybanana”  
(7 results in creative commons)
Tagging behavior

From a set of 52 million photos with 188 million tags, and about 3.7 million unique tags.

Tagging behavior

From a set of 52 million photos with 188 million tags, and about 3.7 million unique tags.

Tagging behavior

From a set of 52 million photos with 188 million tags, and about 3.7 million unique tags.

<table>
<thead>
<tr>
<th>Tags per photo</th>
<th>Photos</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class I</td>
<td>1</td>
</tr>
<tr>
<td>Class II</td>
<td>2 – 3</td>
</tr>
<tr>
<td>Class III</td>
<td>4 – 6</td>
</tr>
<tr>
<td>Class IV</td>
<td>&gt; 6</td>
</tr>
</tbody>
</table>

Other meta-data associated with photos
Notes
Location data

Geo-tagged data

Location info from:

Drag and drop onto map,
Cell tower location,
GPS coordinates.
Social Aspects

• Contacts (friends/family)
• Favorites (bookmarks)
• Comments
• Interestingness
• Groups
Interestingness

Based on user behavior around photos: quantity and quality of meta-data (including tags and comments), number of favorites, number of users who have assigned meta-data, access patterns like click rates and internal and external views, time varying behavior of these attributes.

US patent 20060242139
Groups
Social influence on tag vocabularies

Contact network provides a way for users to easily follow the photos being uploaded by their friends.

This provides a continuous awareness of the photographic activity of their Flickr contacts, and constant exposure to tagging practices.

Do these relationships affect the formation of tag vocabularies?
Social influence on tag vocabularies

Contact network provides a way for users to easily follow the photos being uploaded by their friends.

This provides a continuous awareness of the photographic activity of their Flickr contacts, and constant exposure to tagging practices.

Do these relationships affect the formation of tag vocabularies?
Social influence on tag vocabularies

Contact network provides a way for users to easily follow the photos being uploaded by their friends.

This provides a continuous awareness of the photographic activity of their Flickr contacts, and constant exposure to tagging practices.

Do these relationships affect the formation of tag vocabularies?

2500 users, paired with two other individuals: one randomly chosen from the rest of the set, and the other from their list of contacts.

Calculated overlap in their tag sets (intersection/union).

Random user pairings are much more likely to have a smaller overlap in common tags, while contacts have a higher overall mean.
Social influence

• Shows a relationship between social affiliation and tag vocabulary formation and use even though the photos may be of completely different subject matter.

• This commonality could arise from:
  – Similar descriptive tags (bright, black and white or other photo features).
  – Similar content (photos taken on the same vacation).
  – Similar subjects (co-occurring friends and family).
Flickr: Who is Looking?
Roelof Van Zwol

Who is looking?
  When, why & from where?
Statistics from usage patterns – click patterns
Number of views a photo has generated within 50 days of upload

Why?
  Predict popularity of data items for ranking, retrieval
  Caching strategies for serving & storing items
Their collection

1.83 million public photos uploaded over 60 days

Tracked for 50 days from upload
Viewing statistics

Number of views per photo ordered by #views

Small fraction generates most views
Less than 10% have more than 8 views
About half only viewed once.
Viewing over Time

Average number of views over 50 days

Lots of views on day 1-2, then declines quickly as time goes on
Users quickly discover interesting photos (most popular 10% slice) – 65% of most popular discovered in first 3 hours

Number of views for top slice approximately the same in 3 hours as second slice in 50 days

Most views happen right away. 45% of views of top slice within 48 hours, 82% of 40-50% slice.
Social Dimension

- Groups, contacts, comments, interestingness, favorites

How social the photographer is -> how popular the photo will be – subscription to user’s stream
Number of pools (groups) the photo belongs to influences #views
More viewed photos get more comments