Preservation Trends in the Blogosphere

Banos V. — **Aristotle University**

Baltas N. — **Imperial College**

Manolopoulos Y. — **Aristotle University**

blog forever

Table of Contents

1. Context & definitions
   - Blog, Blogosphere, Preservation, Blog Preservation, Issues, Relevant projects
2. Issues in blog preservation
   - Aggregation, Preservation, Management
3. Approaches to blog preservation
   - Study weblog structures and semantics
   - Define blog digital preservation strategy
   - Software and case studies

Part 1

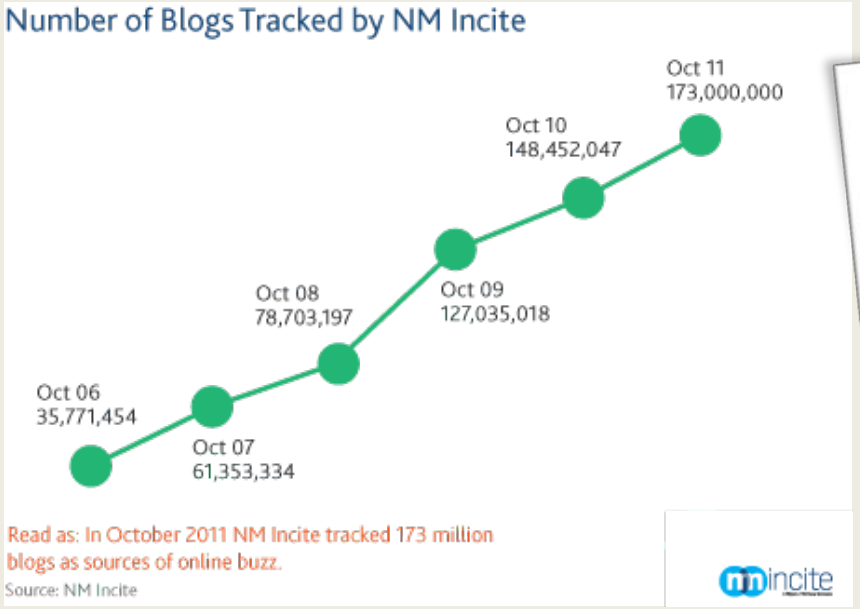
Context and Definitions

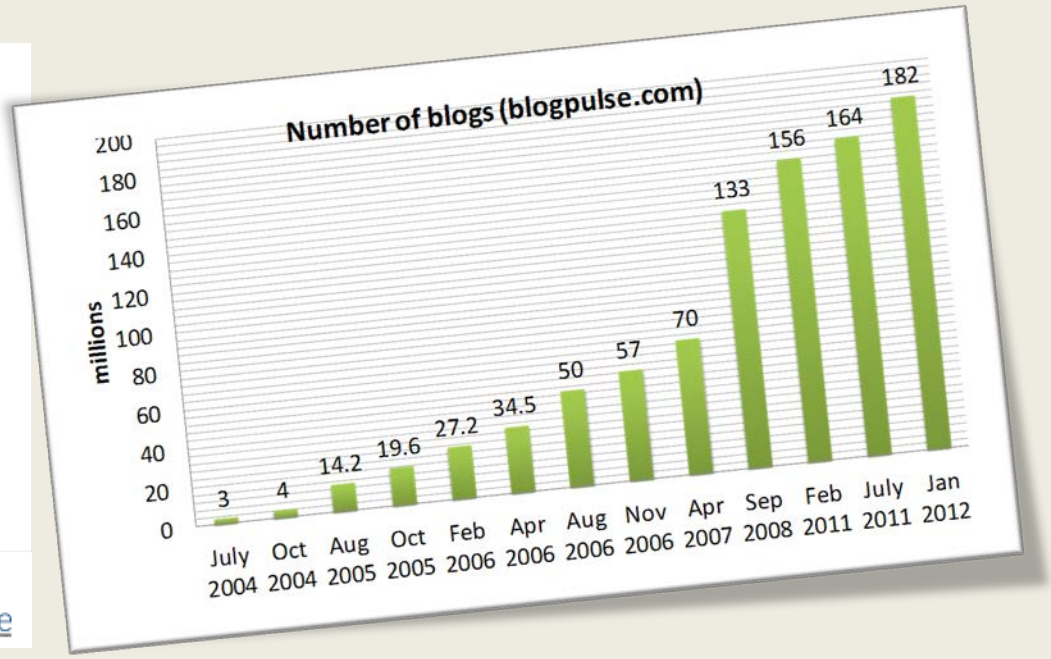
Definitions

- **A blog** is a **web service** that encompasses an **accessible** and widely accepted mechanism for creating, maintaining and automatically distributing **chronologically published material on the Web**, along with the **feedback** and user domain associated with it.
- The **Blogosphere** is the collection of all blogs in the internet and their inter-connections.

State of the Blogosphere

- Blogs have become fairly established as an online communication and web publishing tool.





State of the Blogosphere

- Hundreds of millions of blogs are published about every conceivable subject
- There are many different types of blogs
  - Personal blogs
  - Corporate and organizational blogs
  - By genre (political, health, technology, etc)
  - By media type (vlog, linklog, photoblog, text blog, etc)
  - By device (composed by mobile phone or other computers)
  - Reverse blogs (aggregates of other blogs)

The problem of Blog Preservation

- But despite the fast growth of blogosphere, there is still **no effective solution** for ubiquitous semantic weblog archiving, digital preservation, management and dissemination
- No current Web Archiving effort has ever developed a strategy for effective preservation and meaningful usage of Social Media

Examples for the necessity of blog preservation

- A study (2010) evaluated blog posts for the Iraq war in 2003
  - "Blogs of War: Weblogs as News" (2005) documented 29 blogs on the Iraq war with their names
  - of those 29 blogs,
    - 9 (31%) no longer exist on the Internet,
    - 6 (20.1%) either were abandoned or now deal with a different subject
    - only 13 blogs (45%) were still active
- <u>blogs on major events have already been lost</u>

Examples for the necessity of blog preservation

- Many single owner blogs loose their author through death
  - i.e. Mohammed Nabbous, the most active blogger during the Libyan revolution
  - he was killed in combat and his videoblog still remains today due to his wife who maintains and keeps it alive for the public
- Blogs like these can be lost simply due to the fact that hosting bills are not paid or the blog is deleted due to inactivity.

Digital Preservation

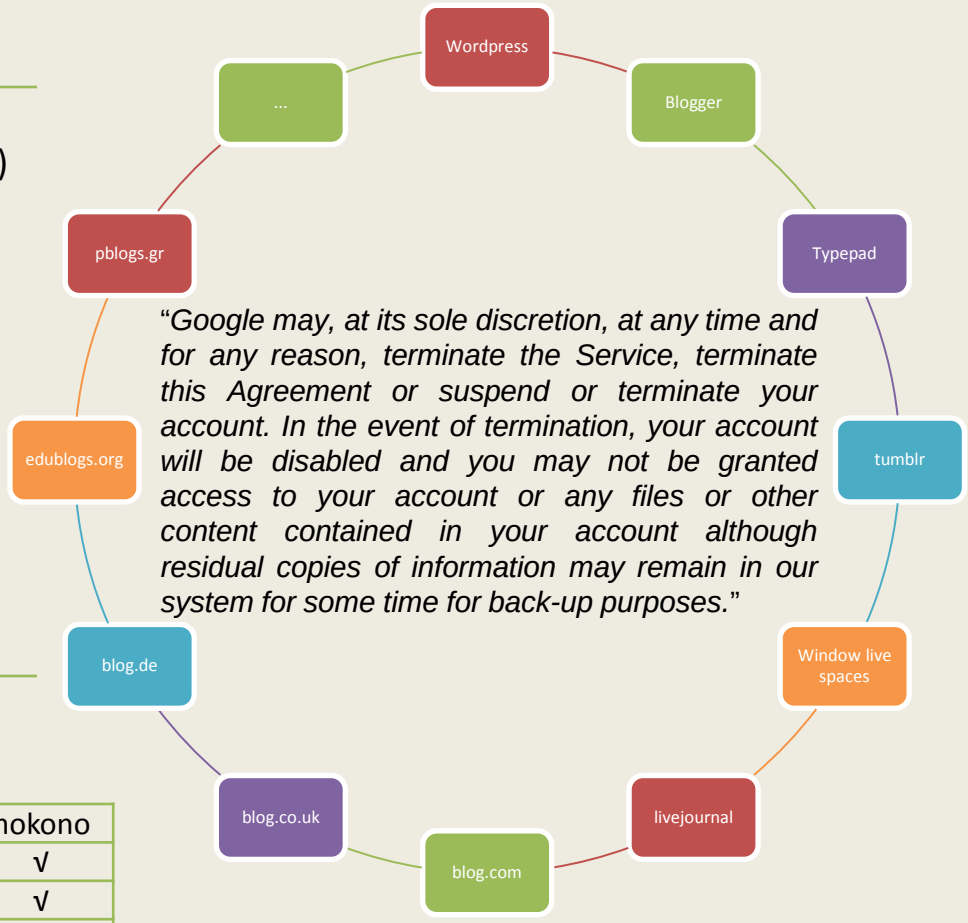
- **Digital preservation** is the set of processes, activities and management of digital information over time to ensure its **long term accessibility**.
- The goal of digital preservation is to preserve materials resulting from digital reformatting, and particularly information that is born-digital with no analog counterpart.
- Because of the relatively short lifecycle of digital information, **preservation is an ongoing process**

Blog Preservation

- The capture, management and preservation of blogs and their resources.
- Aims:
  - availability, viability, fixity, identity, authenticity, understandability, and renderability
- Issues:
  - Frequency of change, quantity and range of resources, integrity of web resources, database driven websites, ownership and DRM, +++

Current state of blog hosting services regarding blog preservation

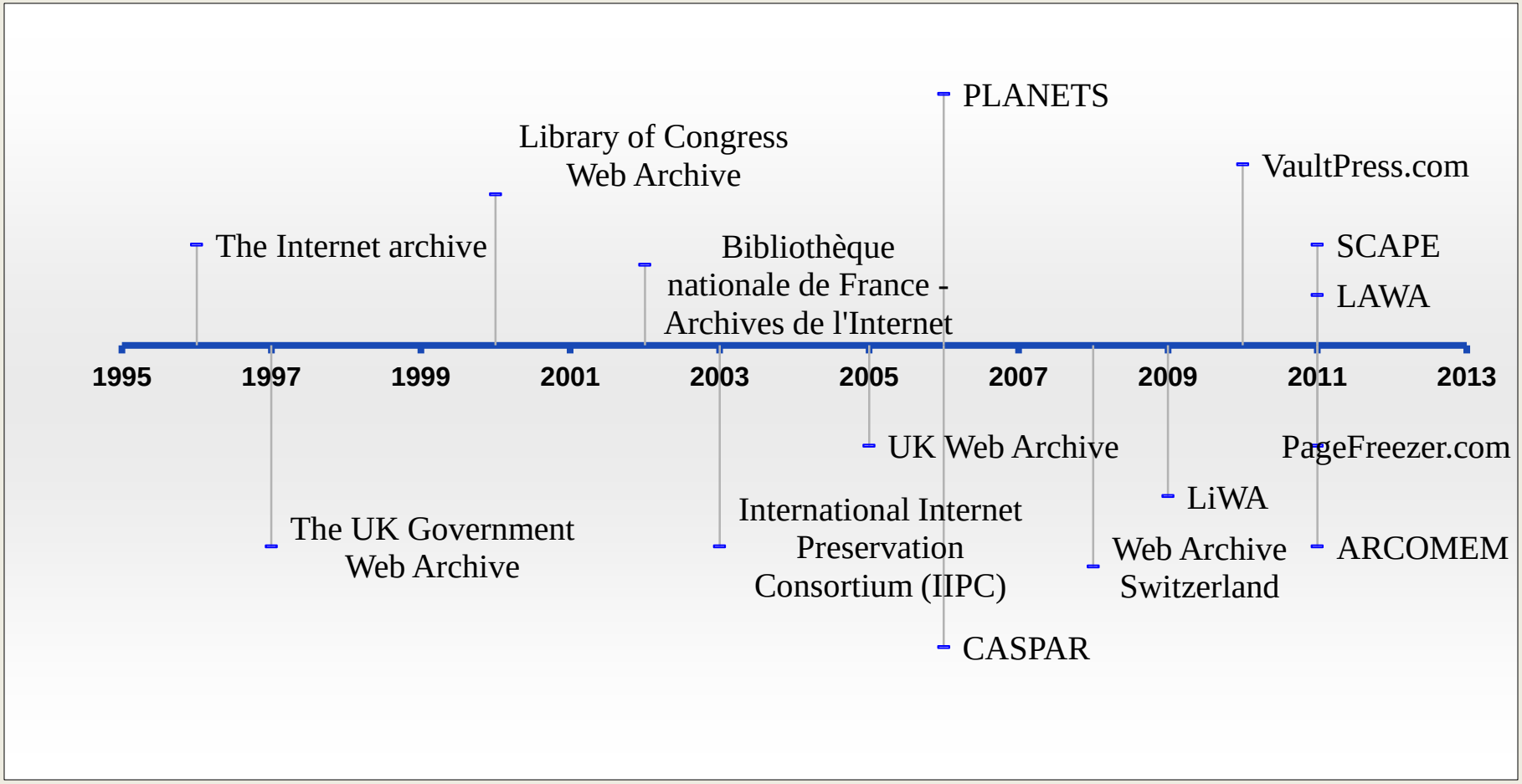
Provider	Based	Global Reach (Unique Users, in Millions)
Blogger	US	61
fc2	JP	120
LiveJournal	US	37
Tumblr	US	110
TypePad	US	28
Wordpress	US	310



"Google may, at its sole discretion, at any time and for any reason, terminate the Service, terminate this Agreement or suspend or terminate your account. In the event of termination, your account will be disabled and you may not be granted access to your account or any files or other content contained in your account although residual copies of information may remain in our system for some time for back-up purposes."

	blogger	wordpress	tumblr	typepad	mokono
Automatic archives	√		√	√	√
Manual backup	√	√	√	√	√
XML format	√	√	√	√	√
Export files	-	-	-	√	-

Web Preservation Projects and Initiatives



Current Web Archive Services Examples

Organization	Year	Access Methods
Bibliotheca Alexandrina's Internet Archive, Egypt	1996	URL Search
Bibliothèque nationale de France - Archives de l'Internet	2002	URL Search, Keyword Search, Full-Text Search, Topical Collections
Government of Canada Web Archive	2005	URL Search, Keyword Search, Alphabetic Browsing, Full-Text Search
The Internet Archive (International)	1996	URL Search, Topical Collections
Library of Congress Web Archive, USA	2000	URL Search, Alphabetic Browsing, Subject Browsing, Topical Collections
National Library of Korea	2005	URL Search, Keyword Search, Subject Browsing
PANDORA Australia's Web Archive	1996	URL Search, Keyword Search, Alphabetic Browsing, Full-Text Search, Subject Browsing
UK Web Archive	2005	URL Search, Alphabetic Browsing, Full-Text Search, Subject Browsing, Topical Collections
Web Archive Switzerland	2008	URL Search, Keyword Search, Full-Text Search, Subject Browsing, Topical Collections

Relevant EU Projects

- **LiWA** – Living Web Archives
- **ARCOMEM** – collect-all ARchives to COmmunity MEMories
- **SCAPE** – SCAlable Preservation Environments
- **LAWA** – Longitudinal Analytics of Web Archive Data
- **BlogForever -** create a software platform capable of aggregating, preserving, managing and disseminating blogs

LiWA - Living Web Archives

- Extend the current state of the art and develop the next generation of Web content capture, preservation, analysis, and enrichment services to **improve fidelity, coherence,** and **interpretability of web archives**
- developing methods which improve archive fidelity
- developing methods for improved archive coherence and interpretability
- focusing on **audiovisual streams** and **social web** content
- http://liwa-project.eu/

ARCOMEM - collect-all ARchives to COmmunity MEMories

- innovative models and tools for Social Web driven content appraisal and selection, and intelligent content acquisition
- novel methods for Social Web analysis, Web crawling and mining, **event and topic detection** and consolidation, and multimedia content mining
- reusable components for archive enrichment and contextualization
- two complementary example applications, the first for **media-related Web archives** and the second for **political archives**
- a standards-oriented ARCOMEM demonstration system
- http://www.arcomem.eu

SCAPE – SCAlable Preservation Environments

- develop **scalable services** for planning and execution of preservation strategies on a **web scale**
- develop infrastructure and tools for scalable preservation actions;
- Provide a framework for automated, quality-assured preservation workflows
- integrate these components with a policy-based preservation planning and watch system.
- validated within three large-scale testbeds from diverse application areas.
- http://www.scape-project.eu/

LAWA– Longitudinal Analytics of Web Archive Data

- The LAWA project on Longitudinal Analytics of Web Archive data will build an Internet-based experimental testbed for **large-scale data analytics**.
- Develop a sustainable infra-structure, **scalable methods,** and easily usable software tools for aggregating, querying, and analyzing heterogeneous data at **Internet scale.**
- Particular emphasis will be given to longitudinal data analysis along the time dimension for Web data that has been crawled over extended time periods.
- http://www.lawa-project.eu

BlogForever Partners

	Name	Short name	Country
Academic Partners	Aristotle University of Thessaloniki	AUTH	Greece
	CERN, The European Organization for Nuclear Research	CERN	Switzerland
	University of London	UL	UK
	University of Glasgow	UG	UK
	University of Warwick	UW	UK
	Technical University Berlin	TUB	Germany
Business Partners	Software Research and Development Consultancy	SRDC	Turkey
	Tero Research & Technology Consulting	Tero	Greece
	CyberWatcher	CW	Norway
	ALTEC Software Development S.A.	ALTEC	Greece
Blog Content Providers	Phaistos Networks S.A.	Phaistos	Greece
	Mokono GmbH	Mokono	Germany

The BlogForever Platform

blog forever

BlogForever will create a software platform capable of aggregating, preserving, managing and disseminating blogs.

Key Objectives

BlogForever will develop robust **digital preservation, management** and **dissemination** facilities for weblogs.

These facilities will be able to **capture** the dynamic and continuously evolving nature of weblogs, their **network** and **social structure**, and the exchange of concepts and ideas that they foster; pieces of information omitted by current Web Archiving methods and solutions.

Scientific Objectives

- Study weblog structure and semantics
- Define a robust digital preservation policy for weblogs
- Implement a weblog digital repository
- Implement specific case studies

The BlogForever Platform

The BlogForever Platform

Impact

The final output of BlogForever will be a simple weblog digital archiving solution that any user, user group or institution could use to preserve their weblog(s) and ensure their authenticity, integrity, completeness, usability, and long term accessibility as a valuable cultural, social, and intellectual resource.

A multitude of parties will benefit from the project (Bloggers, Universities, Libraries & Information Centers, Museums, Education, Research, Business)

Advances to the state of the art

- Definition of a generic data model for weblog metadata and semantics
- Weblog digital preservation strategies
- Weblog spider
- Weblog digital repository web application

Part 2

Issues in blog preservation

Blog Content Aggregation Issues

- Web content aggregation scheduling
  - Weblogs are extremely volatile
  - Not considering web site updates, only perform periodic aggregation
- Web content aggregation performance
  - Brute force techniques, waste computing resources
  - Important components of blogs not identified
- Quality assurance checking
  - No automated archive checking
  - Not identify important information (metadata, microformats, specific blog elements of importance)

Blog Content Preservation Issues

- Current web preservation initiatives are geared towards aggregating and preserving **html pages** and not **information entities** (posts, comments, authors, metadata, dates, pingbacks, etc)
- **Current web archiving efforts** disregard the preservation of **Social Networks** and **interrelations** between the archived content (meme-effect)
- **Current web archives** cannot identify **topics**, **subjects** or **events** (monolithic). There is no generic web archiving solution capable to implement arbitrary subjects and topic hierarchies.

Blog Archive Management Issues

- Regardless of the way a weblog is archived, current solutions do not provide users with meaningful management features of the stored information.
- Current solutions completely disregard the social aspect and interrelations of weblogs or other social media.
- Even browsing the preserved Blogosphere through current Web Archiving solutions, like Internet Archive remains a desired if not impossible task.
- No meaningful information - **need of ontologies**

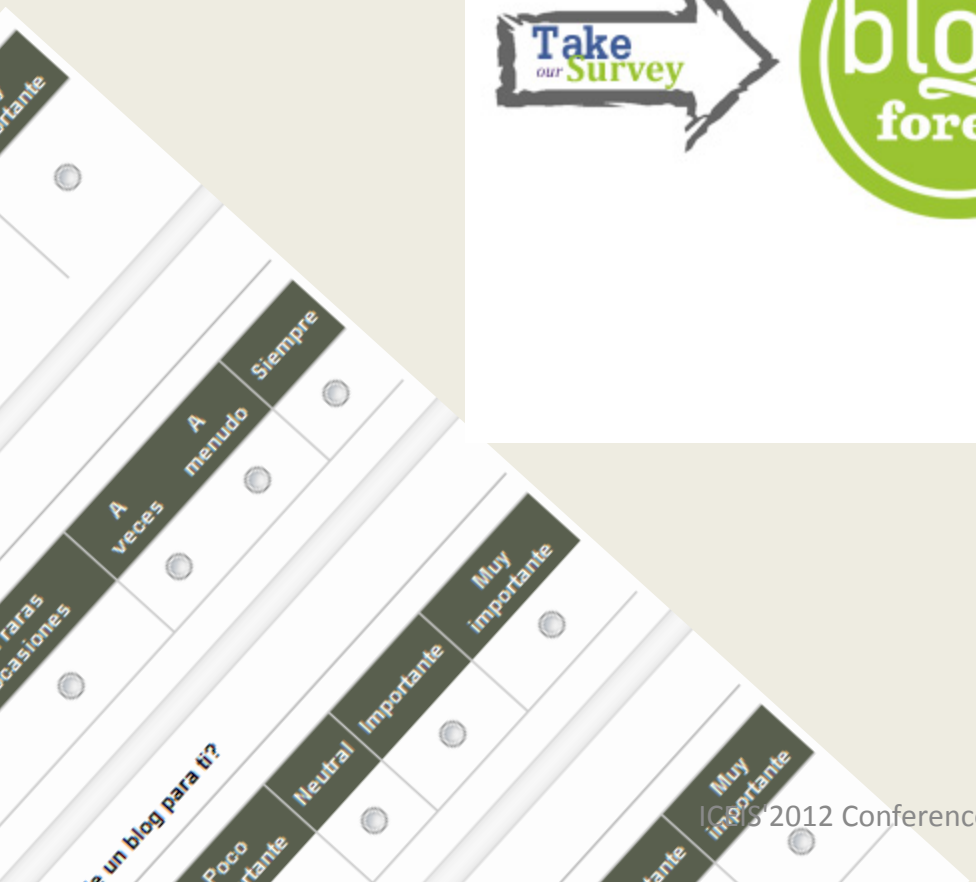
Part 3

Directions towards robust blog preservation

BlogForever

- Study weblog structure and semantics
- Define a robust digital preservation strategy for weblogs
- Implement a weblog digital repository
- Implement specific case studies
- http://www.blogforever.eu

BlogForever Survey



BlogForever Survey Aims

Main aims include the study of:

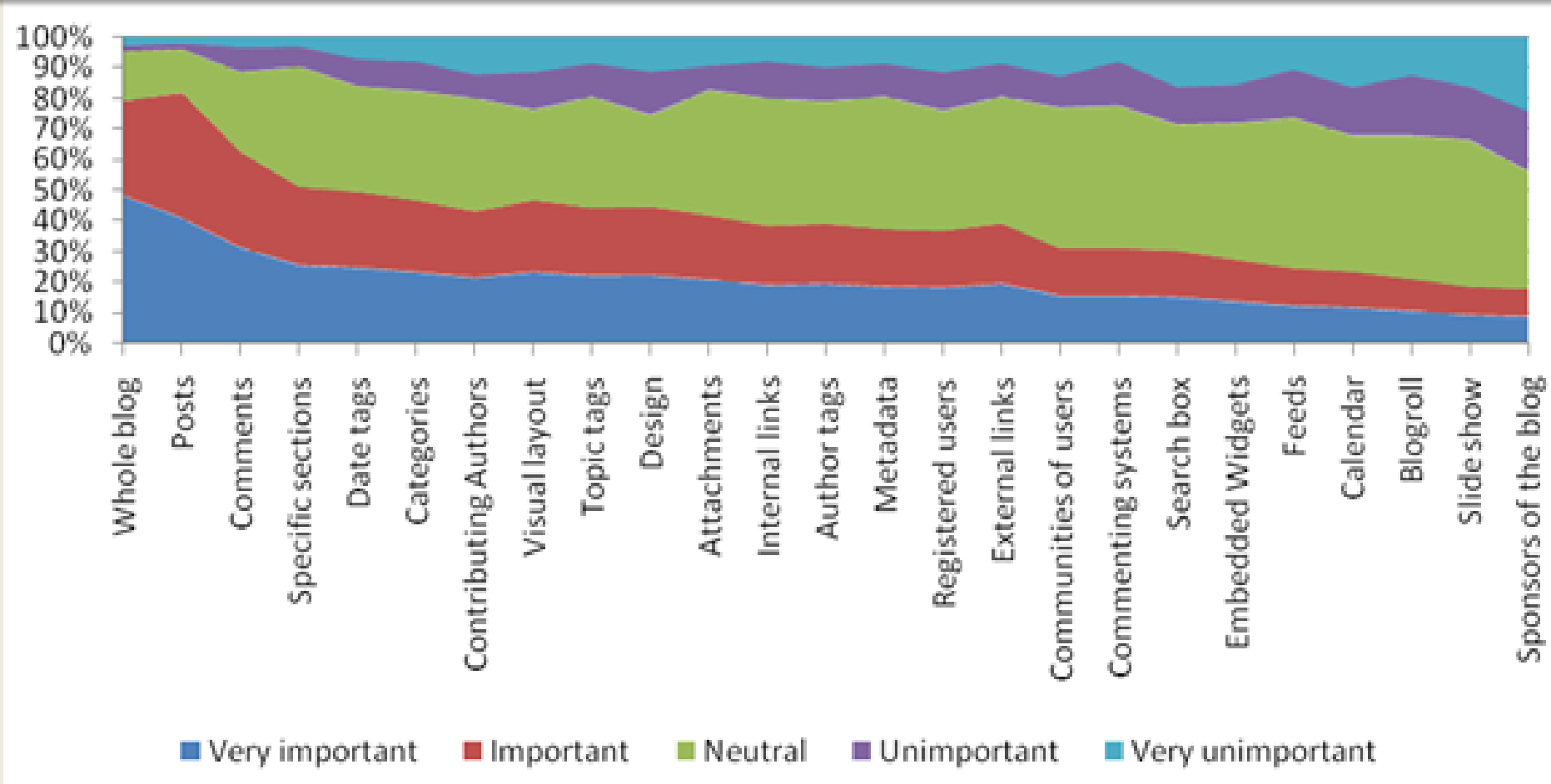
- ✓ *Common blog **authoring practices***
- ✓ ***Patterns** in blog structure*
- ✓ *Blog **lifecycles** and ranking*
- ✓ *Blog author **intention** to contribute to a blog archive*
- ✓ ***factors for search** strategies in a blog archive*
- ✓ *Common blog **preservation aspects***

BlogForever Survey Responders

Responders:

- ✓ *Readers: 517*
- ✓ *Authors: 428*

Importance of Preservation of Blog Elements



BlogForever Survey – General Insights

- ✓ *Majority of authors (90%) never use external services to preserve their blog, **relying on their blog provider**.*
- ✓ *Using a **blog archive is welcome** but authors think blog providers are the leading bodies for this role at the moment.*
- ✓ *Motivations for maintaining blogs were primarily **personal** (80%) for sharing information and promoting discussion topics.*
- ✓ *Perceived importance of **rich media**, links and citations*
- ✓ *Wide variation in the use of media objects*
- ✓ *Primary use: **textual content** (98%) in blogs*
- ✓ *Frequent use of **photographs** and moving images (40%-83%)*

BlogForever Technology Survey

Main goal was to study large number of blogs and evaluate the use of:

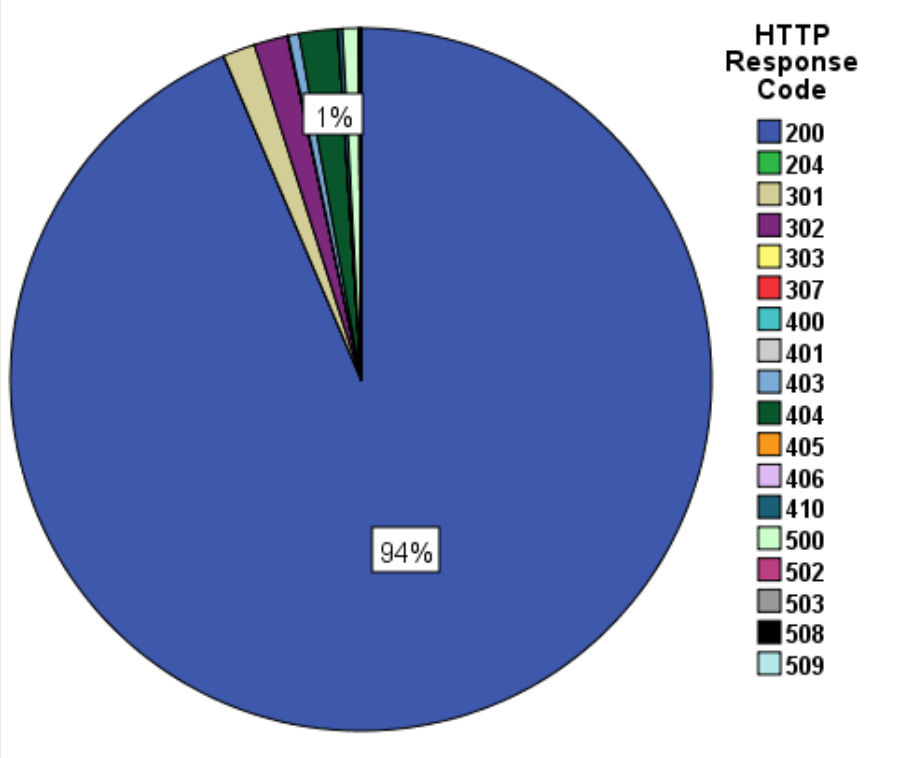
- ✓ *Third-Party Libraries*
- ✓ *External Services*
- ✓ *Semantic Mark-Up*
- ✓ *Metadata*
- ✓ *Web Feeds*
- ✓ *Various Media Formats*

Technology Survey Data Sources

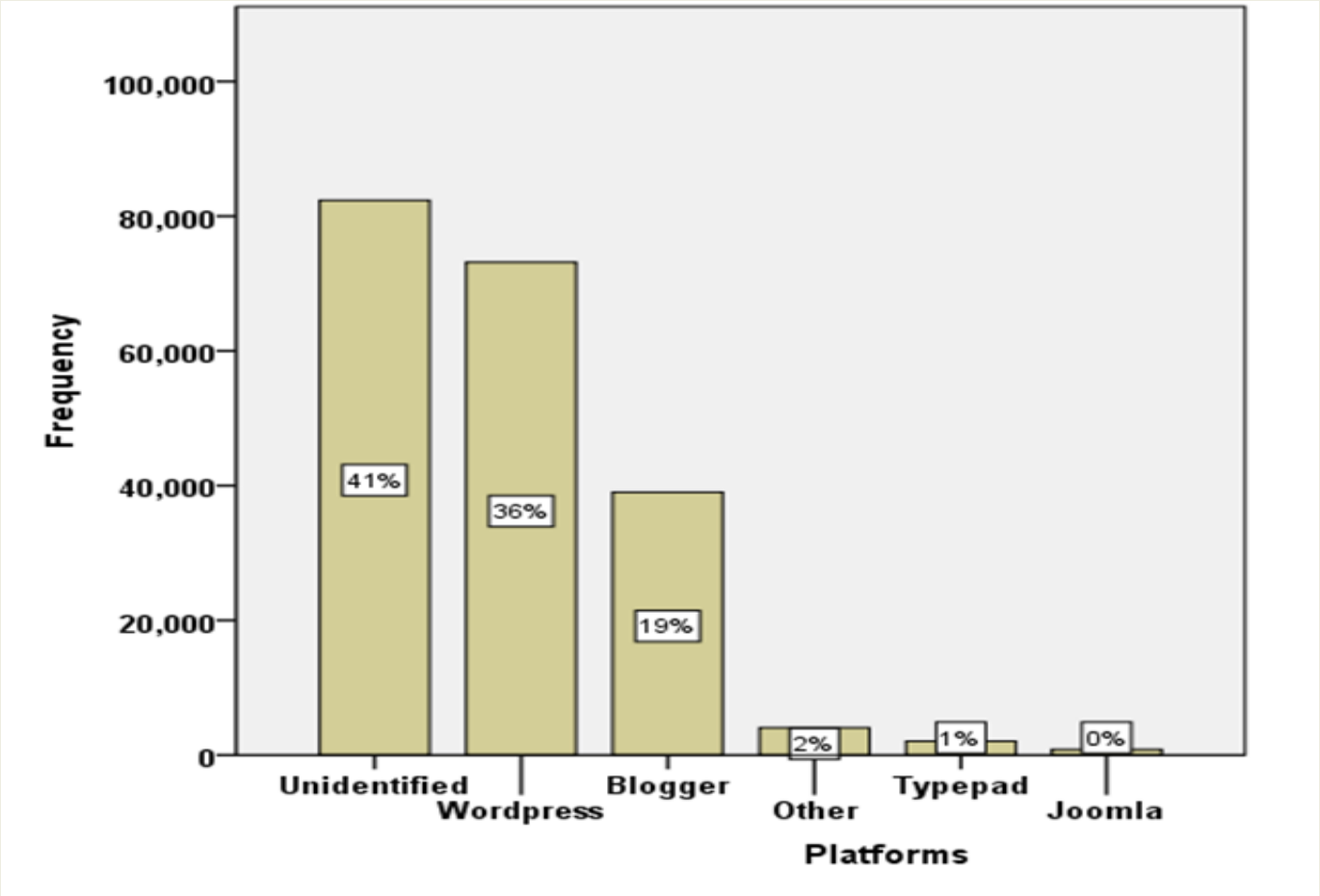
Dataset by weblog.com

Number of included URLs:

- ✓ *Overall total of accessed resources:* ***259,930***
- ✓ *Overall total of valid records:* ***209,830***



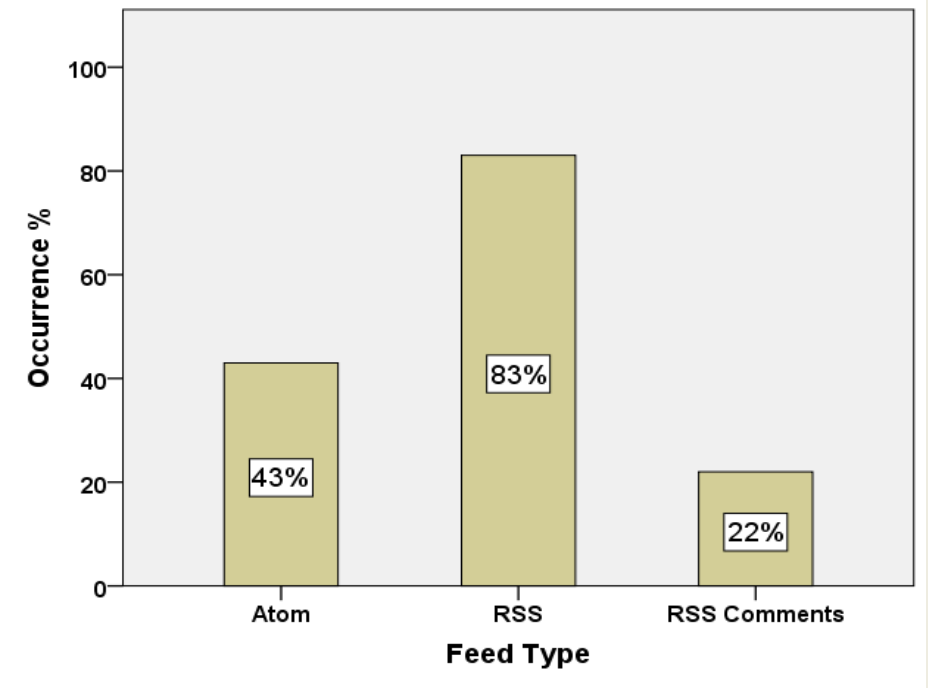
Blog Technology Survey: Blog platforms



Feed type

Frequency of Feeds per total URI (blog):

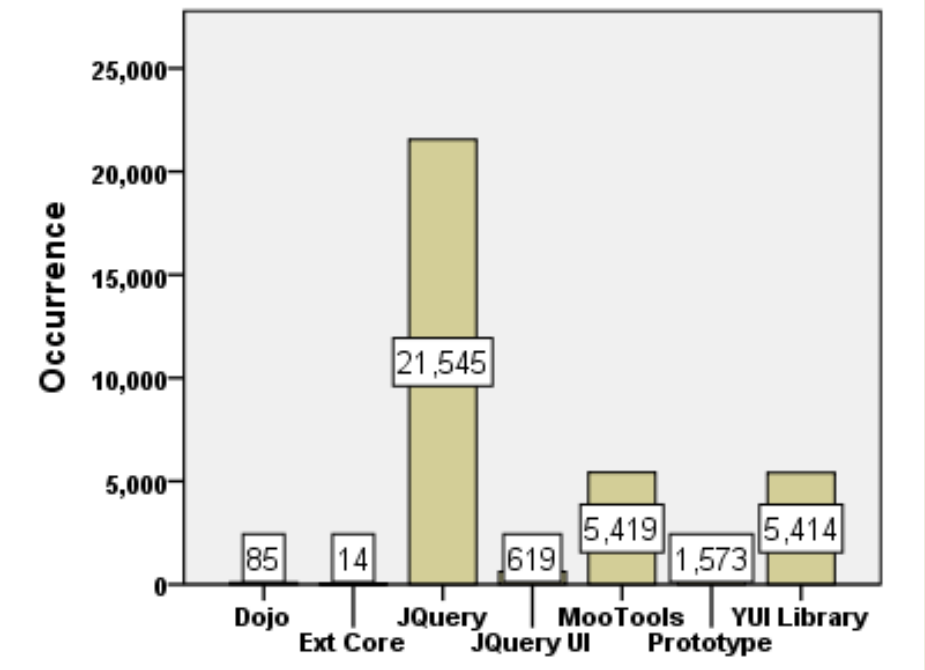
- ✓ *83% - RSS feeds*
- ✓ *43% - Atom feeds ~⊂ RSS*
- ✓ *22% - RSS Comments*

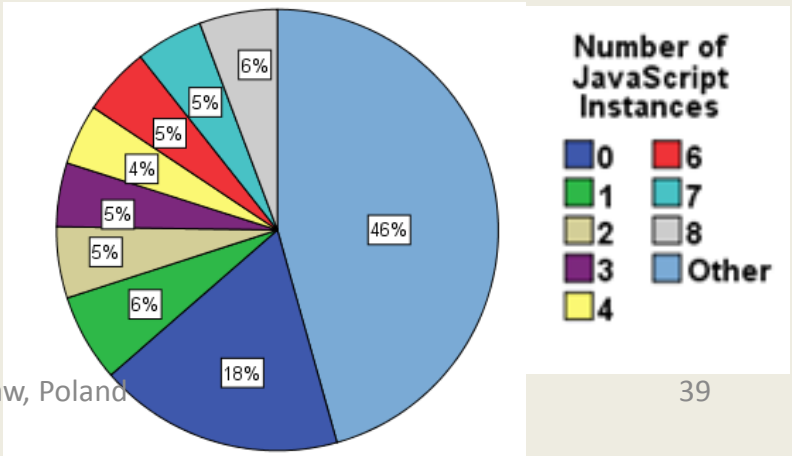


Blog Technology Survey: APIs and Libraries

Most popular Libraries:

- ✓ *Dojo*
- ✓ *Ext Core*
- ✓ *JQuery*
- ✓ *JQuery UI*
- ✓ *MooTools*
- ✓ *Prototype*
- ✓ *YUI Library*





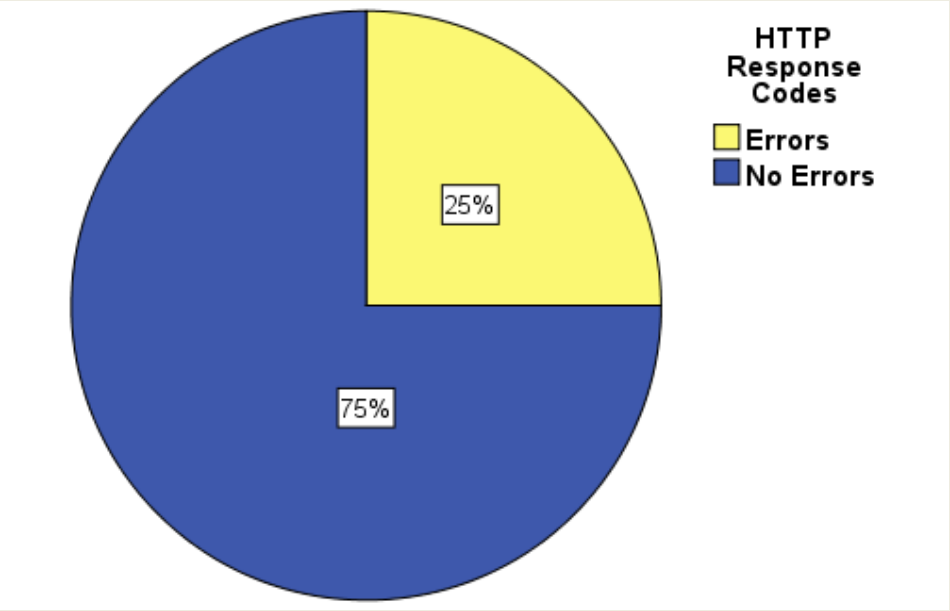
Blog Technology Survey: File Types

File Ext.	Application	Instances
doc	Word Processing	1097
docx	Word Processing	147
odt	Word Processing	51
pdf	Word Processing	13731
txt	Word Processing	641
mp4	Video/Audio	3265
mpeg	Video	36
mpg	Video	613
avi	Video	3265
mov	Video	71
3gpp	Video	1429
xls	Spread Sheet	138
xlsx	Spread Sheet	24

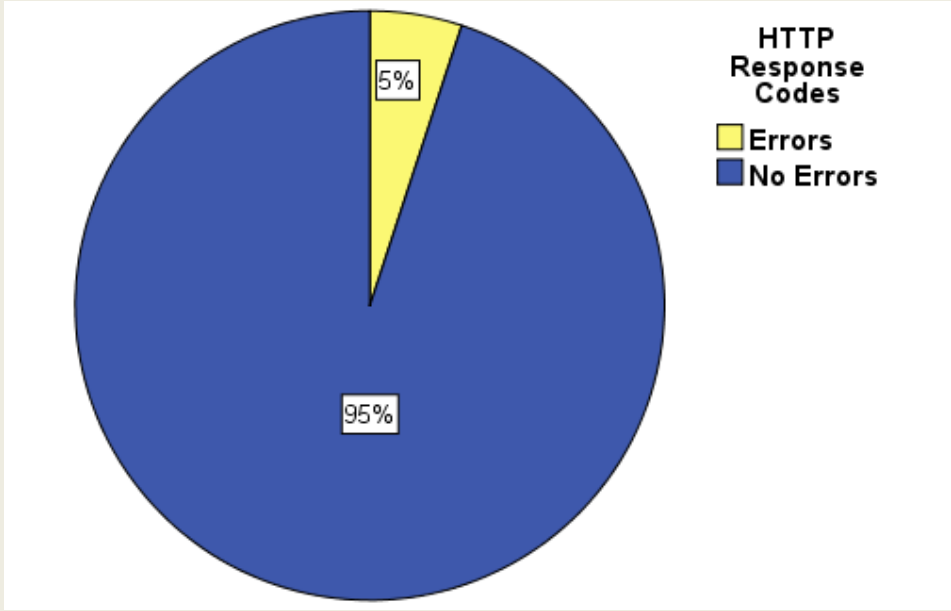
File Ext.	Application	Instances
ods	Spread Sheet	722
ppt	Presentation	67
pptx	Presentation	20
odd	Presentation	618
odf	Math Formulas	63
odg	Graphics	4
mdb	Database	0
ccbd	Database	0
odb	Database	153
vCard	Card	14
mp3	Audio	10231
wav	Audio	13
vrml	3D	0

Comparing blogs and web in general Errors

The Web

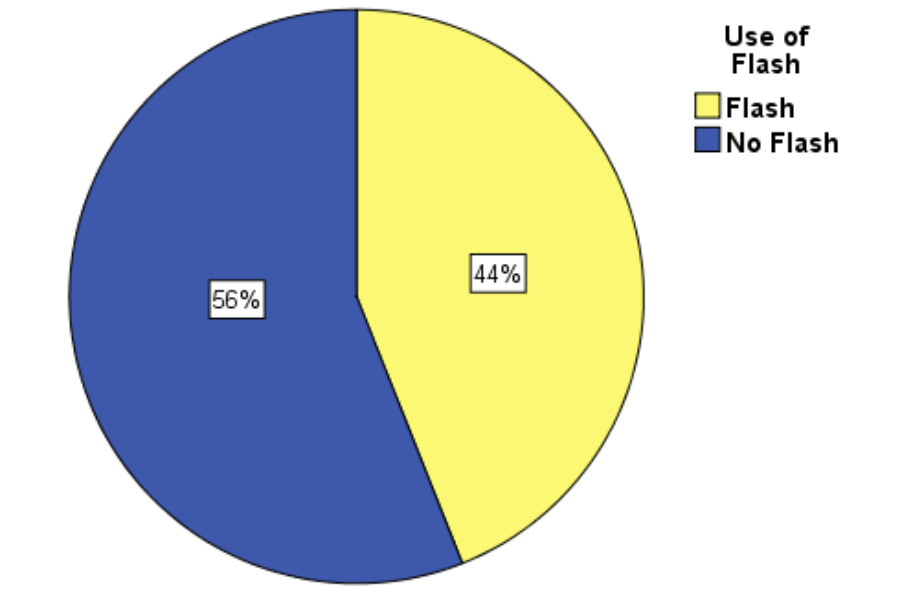


The Blogosphere

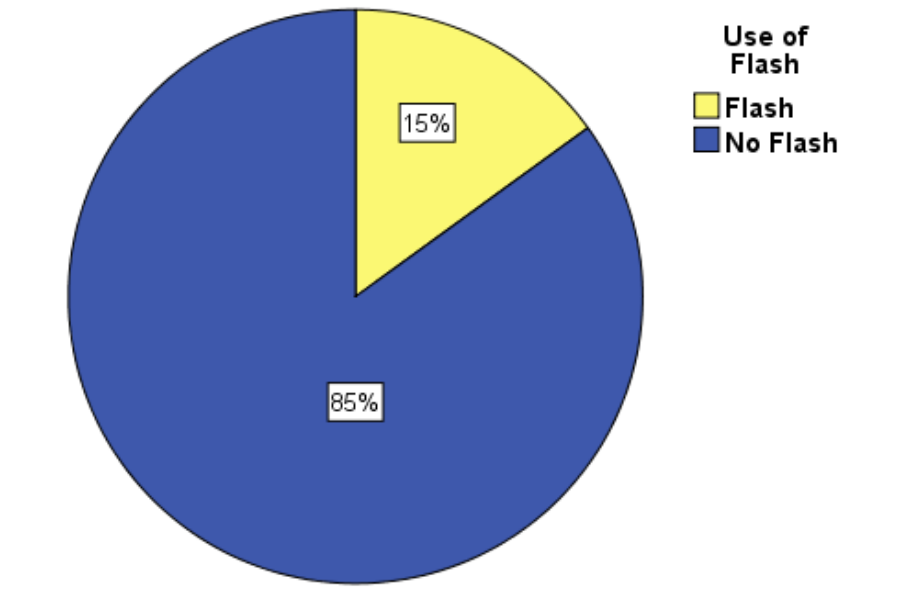


Comparing blogs and web in general
Use of Flash

The Web

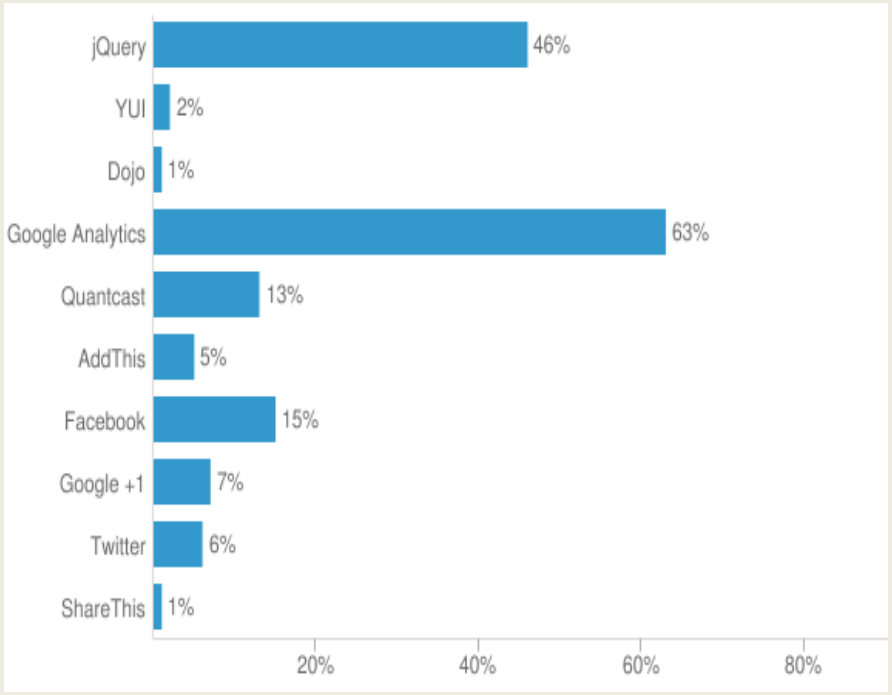


The Blogosphere

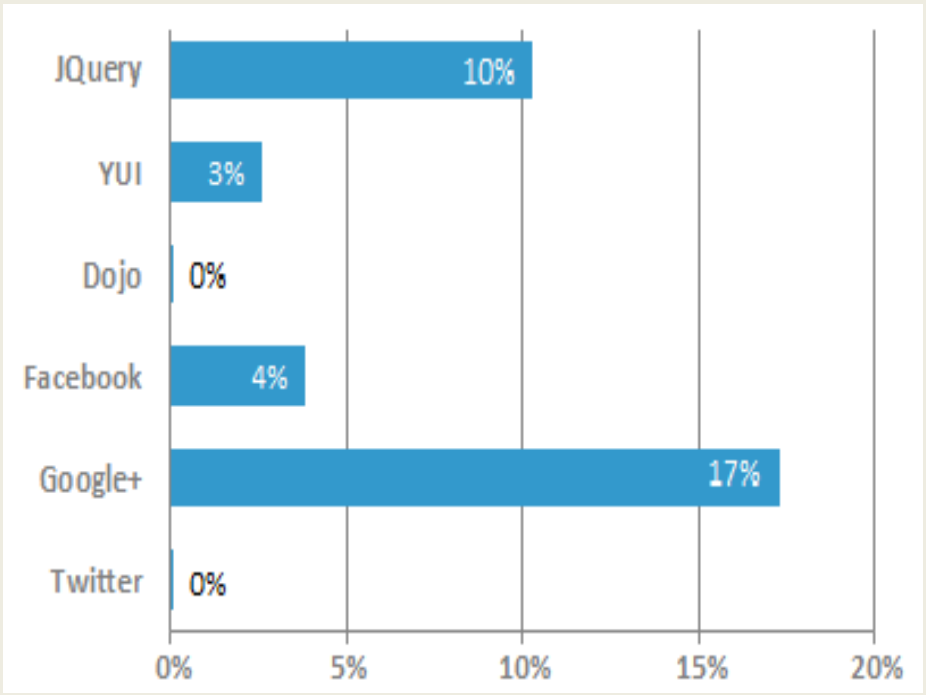


Comparing blogs and web in general Common JS Libraries

The Web

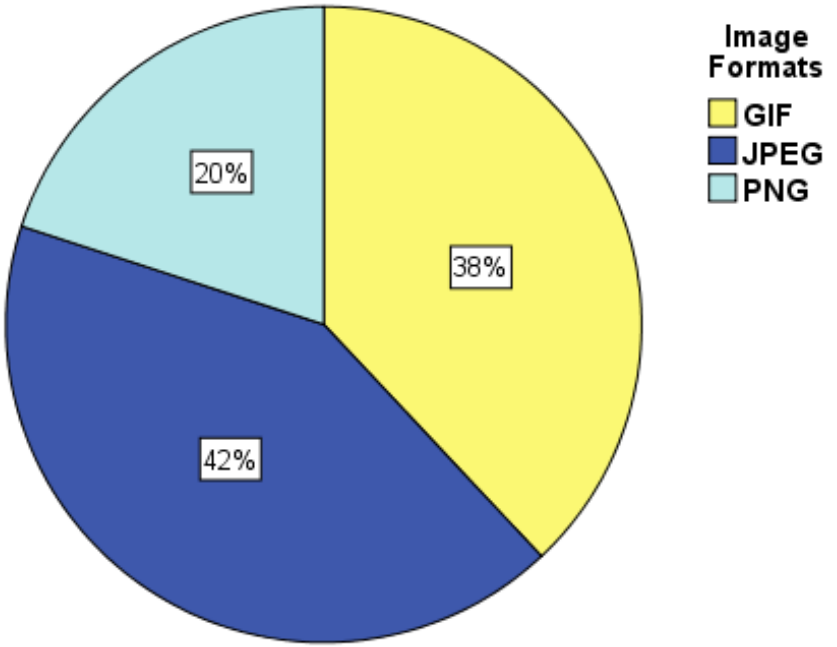


The Blogosphere

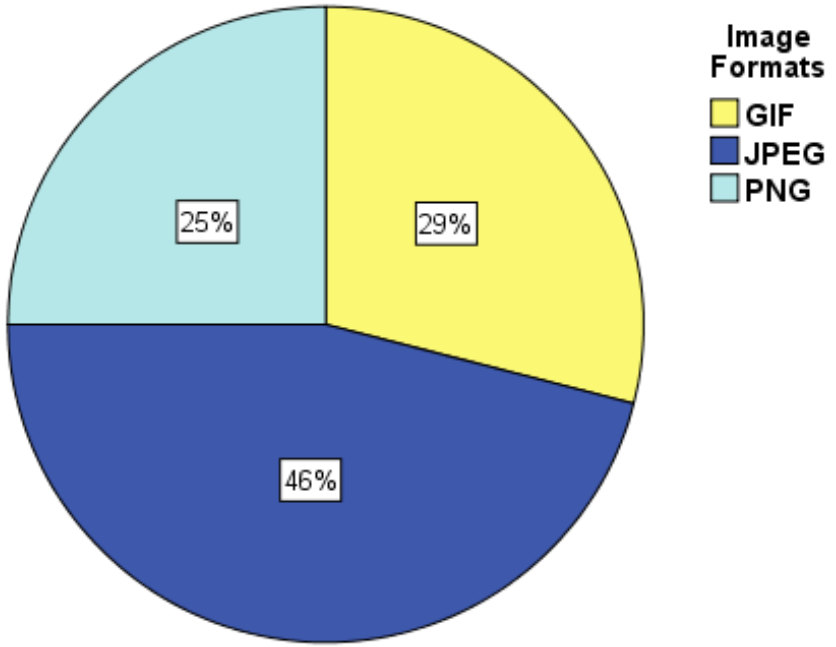


Comparing blogs and web in general
Image Formats

The Web



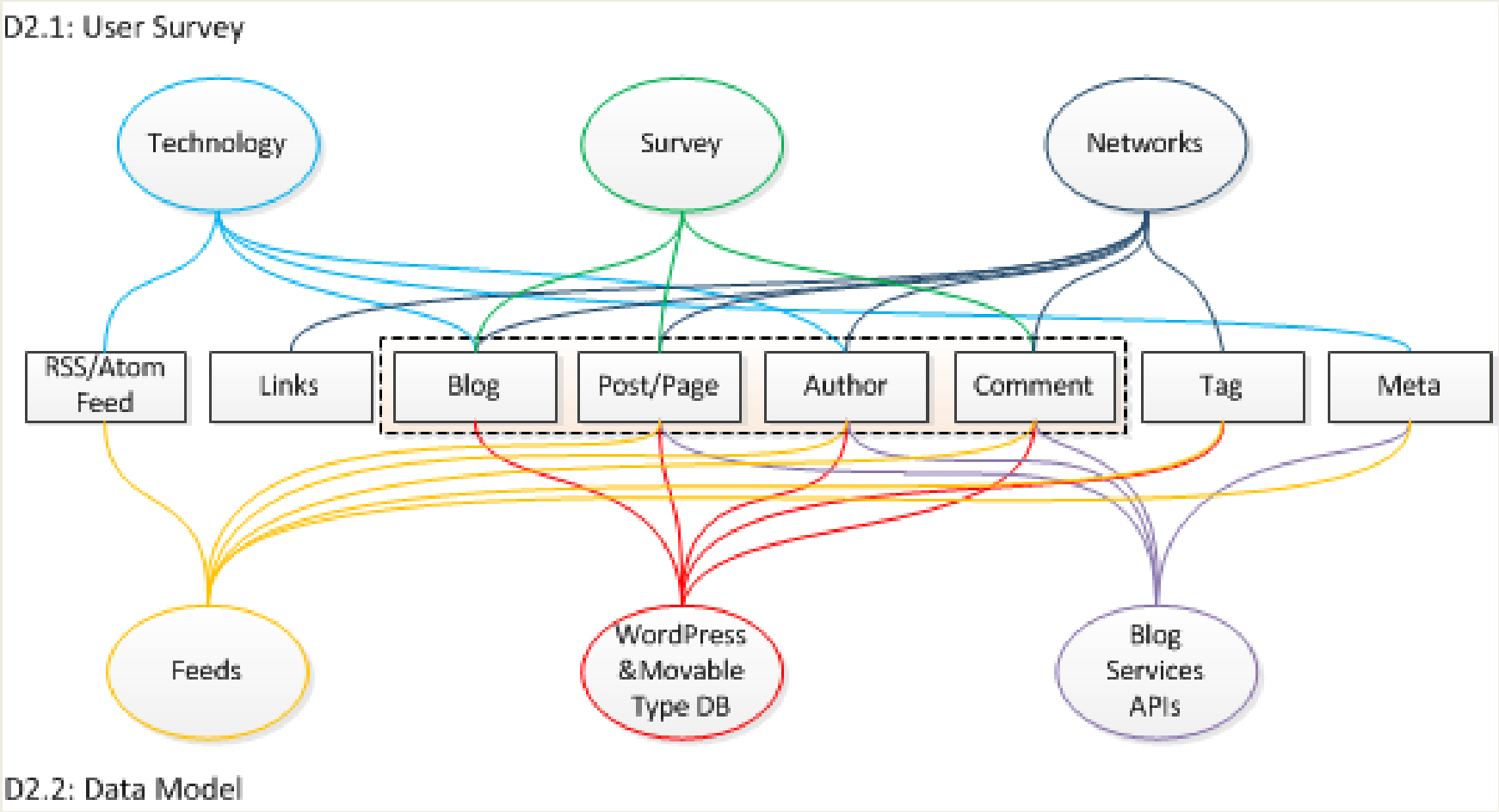
The Blogosphere

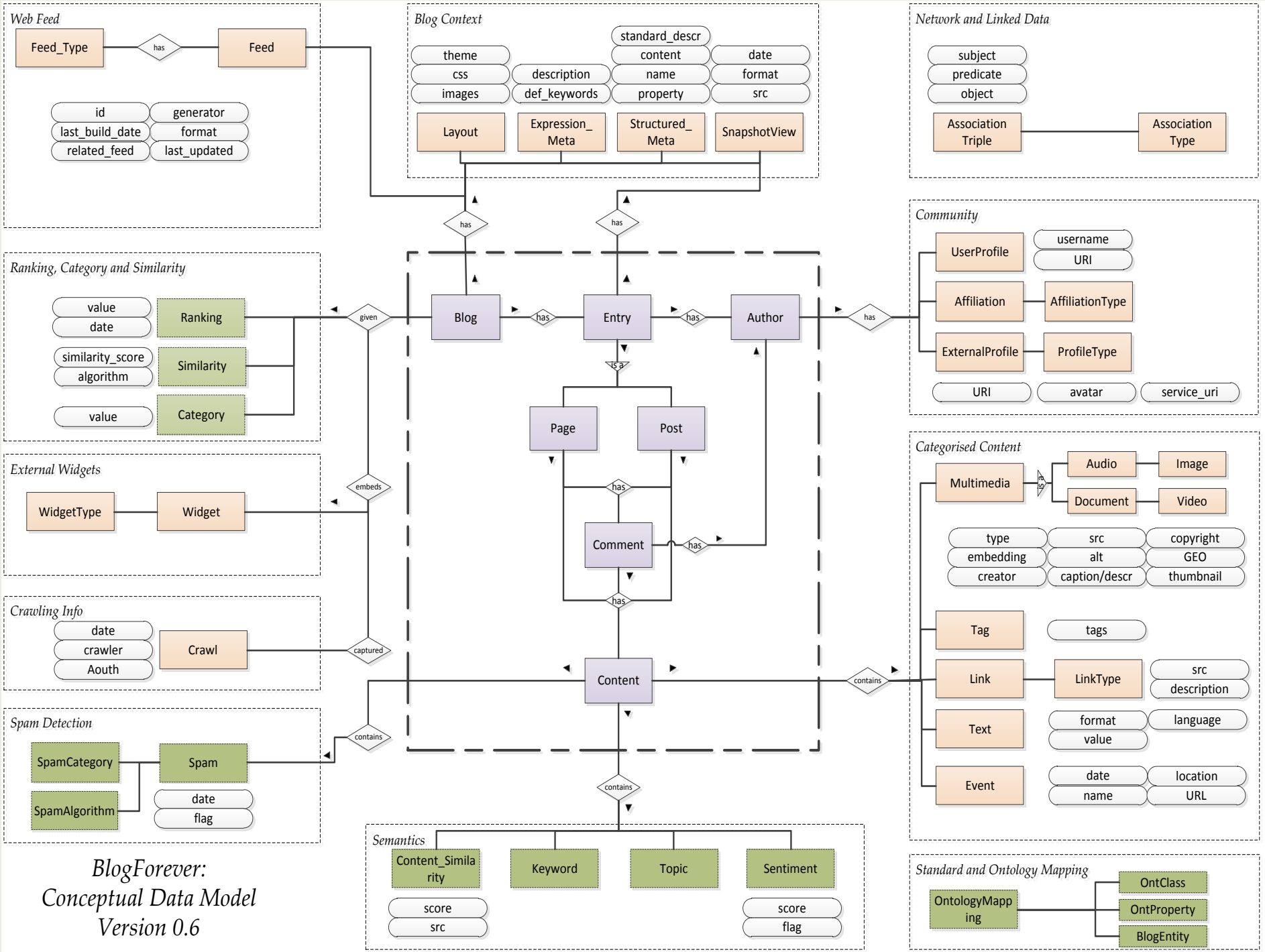


Blog Technology Survey Conclusions

- *469 platforms (with the dominant WordPress & Blogger)*
- *Wide range of encoding standards*
- *RSS/Atom feeds, CSS and JavaScript are among most widely used technologies*
- *Use of images and their formats are similar to their general use on the Web*
- *The use of Dublin Core, Open Graph, FOAF and SIOC is not widespread*
- *Quick adoption of Google+ and evidence of some integration of Facebook and Twitter*

Blog modeling and semantics



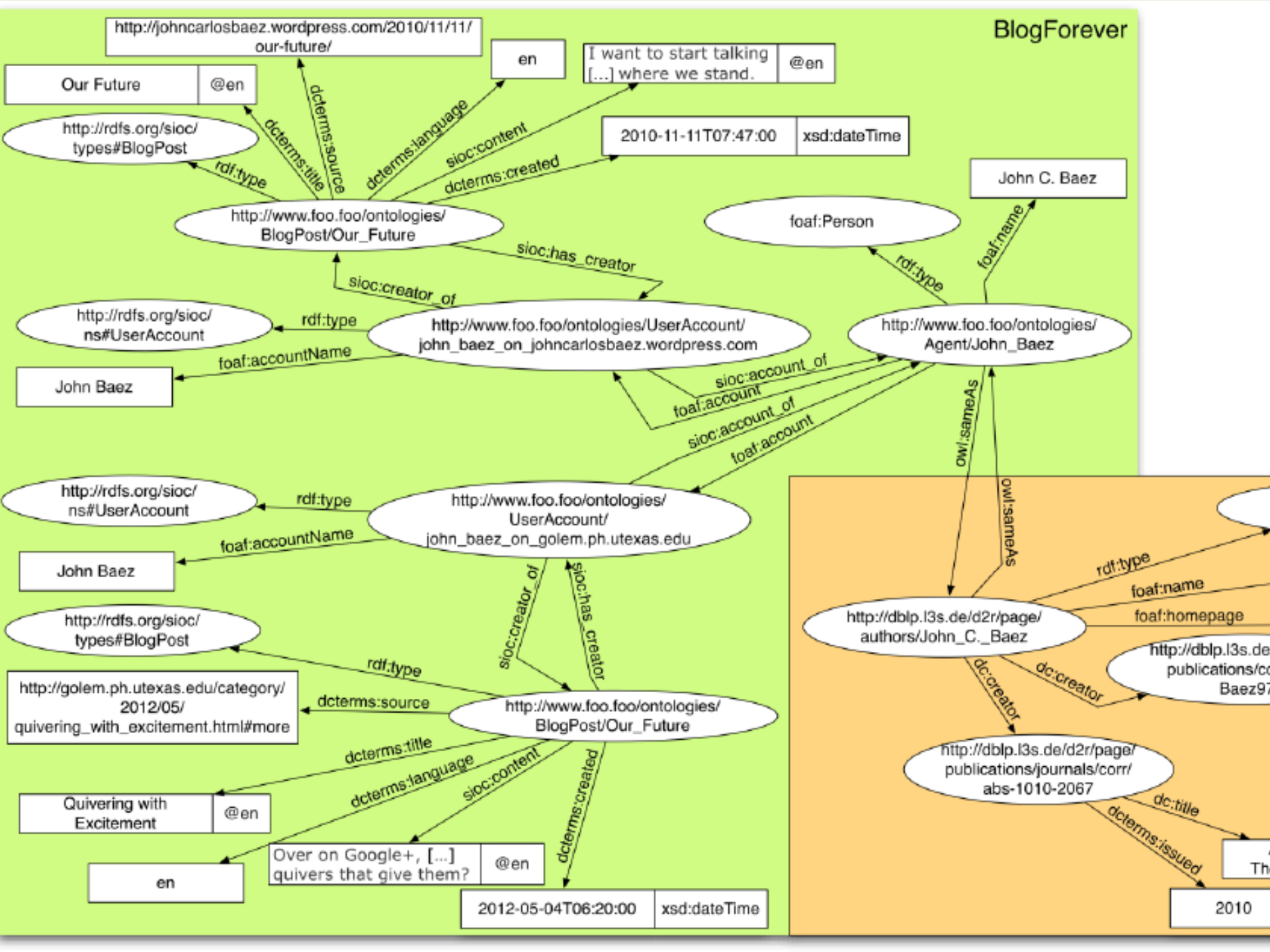

Web Feed
Feed_Type
has
Feed
id
generator
last_build_date
format
related_feed
last_updated
Blog Context
theme
css
images
description
def_keywords
standard_descr
content
name
property
date
format
src
Layout
Expression_ Meta
Structured_ Meta
SnapshotView
Network and Linked Data
subject
predicate
object
Association Triple
Association Type
Community
UserProfile
username
URI
Affiliation
AffiliationType
ExternalProfile
ProfileType
URI
avatar
service_uri
Ranking, Category and Similarity
value
date
Ranking
similarity_score
algorithm
Similarity
value
Category
given
Blog
has
Entry
has
Author
has
is a
Page
Post
External Widgets
WidgetType
Widget
embeds
has
Comment
has
has
Crawling Info
date
crawler
Aouth
Crawl
captured
Content
contains
Categorised Content
Multimedia
is a
Audio
Image
Document
Video
type
embedding
creator
src
alt
caption/descr
copyright
GEO
thumbnail
Tag
tags
Link
LinkType
src
description
Text
format
value
language
Event
date
name
location
URL
Spam Detection
SpamCategory
SpamAlgorithm
Spam
date
flag
contains
contains
Semantics
Content_Similarity
score
src
Keyword
Topic
Sentiment
score
flag
Standard and Ontology Mapping
OntologyMapping
OntClass
OntProperty
BlogEntity
BlogForever:
Conceptual Data Model
Version 0.6

Future Work: Analyse Archived Blogs

- Analyse blog archives to gain a better understanding of the content and provide new services:
  - Use **Linked Open Data** to link archived blog content with other web content
  - Apply **Semantic Extension of Tags** to understand them better and reuse them for multiple purposes.
- In any case, use **Ontologies** to interpret and reason with information.

Linked Open Data

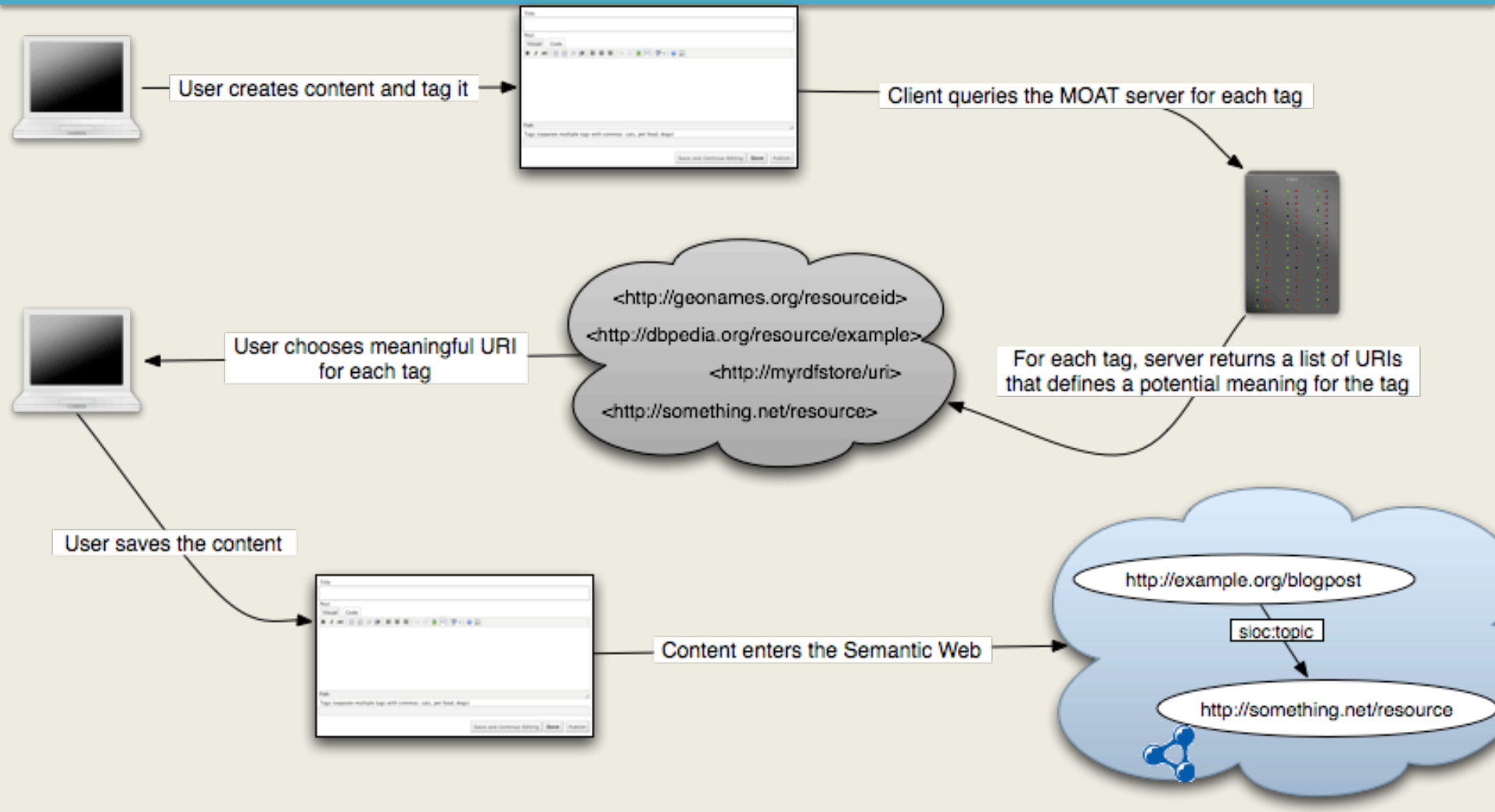
- Linked Open Data can help us link archived blog content with other web content.
- Improve interoperability by using common vocabularies:
  - Friend of a Friend (FOAF)
  - Dublin Core (DC)
  - Semantically Interlinked Online Communities (SIOC)
- Example: Link archived blog content to the description of scientific publications in DBLP


BlogForever
http://johncarlosbaez.wordpress.com/2010/11/11/our-future/
en
I want to start talking [...] where we stand.
@en
Our Future
@en
dcterms:source
dcterms:language
sioc:content
http://rdfs.org/sioc/types#BlogPost
dcterms:title
2010-11-11T07:47:00
xsd:dateTime
rdf:type
dcterms:created
http://www.foo.foo/ontologies/BlogPost/Our_Future
John C. Baez
foaf:Person
foaf:name
rdf:type
sioc:has_creator
sioc:creator_of
http://rdfs.org/sioc/ns#UserAccount
rdf:type
http://www.foo.foo/ontologies/UserAccount/john_baez_on_johncarlosbaez.wordpress.com
http://www.foo.foo/ontologies/Agent/John_Baez
foaf:accountName
John Baez
sioc:account_of
foaf:account
sioc:account_of
foaf:account
owl:sameAs
owl:sameAs
http://rdfs.org/sioc/ns#UserAccount
rdf:type
http://www.foo.foo/ontologies/UserAccount/john_baez_on_golem.ph.utexas.edu
foaf:accountName
John Baez
sioc:creator_of
sioc:has_creator
rdf:type
foaf:name
foaf:homepage
http://dblp.l3s.de/d2r/page/authors/John_C._Baez
http://rdfs.org/sioc/types#BlogPost
rdf:type
http://golem.ph.utexas.edu/category/2012/05/quivering_with_excitement.html#more
dcterms:source
http://www.foo.foo/ontologies/BlogPost/Our_Future
dc:creator
dc:creator
http://dblp.l3s.de/d2r/page/publications/journals/corr/abs-1010-2067
dcterms:title
dcterms:language
sioc:content
dcterms:created
dc:title
dcterms:issued
Quivering with Excitement
@en
en
Over on Google+, [...] quivers that give them?
@en
2012-05-04T06:20:00
xsd:dateTime
2010

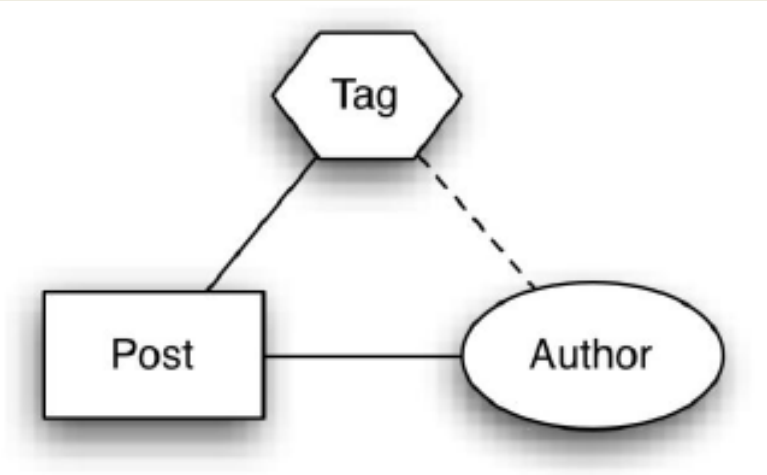
Semantic Extension of Tags

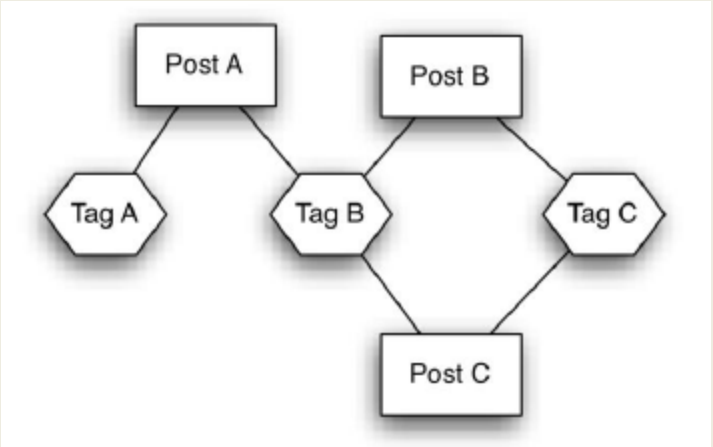
- User generated tags and resulting folksonomies are widespread in social media applications.
- Aims:
  - Identify and expose the meaning of tags using MOAT (Meaning Of A Tag) to resolve various problems
  - Deduce tag relationships to utilize the social aspect of blogs

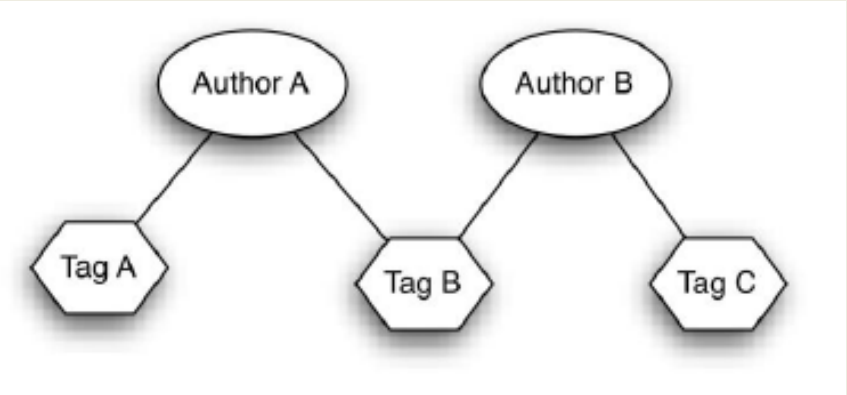
Semantic Extension of Tags: identify and expose their meaning

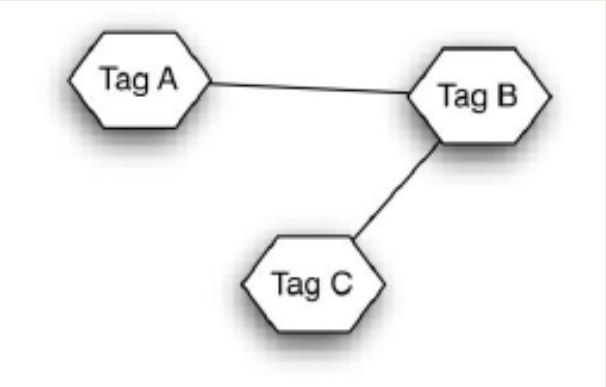


Semantic Extension of Tags: Deduce tag relationships









BlogForever Platform Testing

- 2 academic blogs cases
  - University of Warwick (58 blogs)
  - University of London (70 blogs)
- Multilingual blogs cases
  - 356 blogs in 4 languages with diverse topics
- Multimedia blogs case
  - 1000 Greek blogs with diverse multimedia content
- Large case
  - 500.000 blogs from blog.de

Summary

- Introduction to blog preservation
- Identified a number of open issues on blog preservation
- Presented an outline of current research work on the BlogForever project:
  - Survey
  - Modelling
  - Interoperability
  - Preservation
  - Spam filtering
  - Crawling - Ingestion
  - Software platform
  - Testing using real data

Thank you!

Any Questions?

Visit: http://blogforever.eu to learn more.

The research leading to these results has received funding from the European Commission Framework Programme 7 (FP7), BlogForever project, grant agreement No.269963.