Internet Subscription Services
Introduction

Internet users bundle information from various sources
Explicit information request

- The user initiates requests for all information he is interested in explicitly
- Wanted information has to be retrieved from different network locations
- Problems:
  - High effort for user to get up-to-date information
  - User may miss articles he is interested in
  - Information from different sources are not grouped

Location 1: Online discussions
Location 2: (X)HTML content
Location 3: Media files (e.g. mp3, mpg, ...)
Location 4: (X)HTML content
Subscription services

- Ambitions of subscription services:
  - User is informed about new information automatically
  - Bundling of similar formats from different sources
  - Sophisticated control about which information should be delivered

Publishing

- Web server
- News server
- Media server

Subscribed to a service

different client software
Subscription principles

From the perspective of initiation of information distribution, there exist two principles for realising a subscription service:

1. **Distributor aware subscription**
   - User needs application identifier that makes him accessible in the network permanently (e.g. mail address, Instant Messaging account, ...)
   - Information distributor stores identifier of every user
   - If new information is available, every subscribed user is informed about it automatically and may directly get a copy of it

2. **User client aware subscription**
   - User employs software which he informs about every data source (server) he wants to receive information from
   - Client software accesses all relevant servers at defined times and checks if new information is available (“polling”)
User subscribes to a publishing service by specifying an identifier and some additional information (e.g. which information he wants to receive)

Subscription logic stores this information in a database

If new data is available the publishing service asks for all subscribed users that want to be informed about it

Identifiers are extracted from the database and redirected to the publishing service

Data is delivered to the user; possibly it is temporarily stored on an intermediate server (e.g. mail server)
Mailing lists

- Widespread distributor aware subscription systems are Electronic Mailing lists
- Famous mailing list management software is GNU Mailman which supports use via web interface and email interface
- Commands are simply sent by email to the Mailman server
- Parameters for commands can be included into subject or mail-body
User client aware subscription

1. User subscribes to different subscription services with the use of a subscription client software
2. Addresses of all services are stored
3. Software sends requests for new information to all servers it has been subscribed to in periodic intervals
4. If new information is available, it or its summary is downloaded by the client
5. User is notified about new information
Two subclasses of user client aware subscription systems are:

- **Explicitly designed discussion systems** (e.g. Usenet)
  - Systems are explicitly designed for the purpose of user discussions on the subscription principle
  - Specialised infrastructure and communication protocols
  - Focus on direct user participation

- **Web based subscription systems** (e.g. web feeds)
  - Introduction of additional data representation of World Wide Web content to make it readable for subscription software
  - Web infrastructure and protocol (HTTP) are used for information distribution
Usenet

- Long-established distributed discussion system that is running on thousands of servers; no central control or authority
- Each server defines a set of newsgroups it wants to hold
- Newsgroups are dedicated to specific topics and organised in a hierarchical way (e.g. "comp.infosystems.www.servers" or "news.misc")
- User logs in to a news server and subscribes to newsgroups he is interested in with the use of a special news reader client
- Transmitting articles from server to user client ("reading") or between servers ("transit") which hold the same newsgroups is done by the Network News Transfer Protocol (NNTP)
Usenet message format

- Articles are stored and distributed between servers in a text based format
- Defined in RFC 1036
- Required header elements are "From", "Path", "Newsgroups", "Subject", "Message-ID" and "Date"
- Body has to be separated by a blank line from header

From: harry@mars.dom.com (Harry Hopper)
Path: sysx!serv5!sys3!mars!harry
Newsgroups: news.announce
Subject: Important news notification
Message-ID: <45223423@mars.dom.COM>
Date: Fri, 17 Nov 1970 17:05:42 GMT
Followup-To: news.misc
Expires: Thu, 1 Jan 2030 00:00:00 -0500
Organization: Mars and Son Inc.

This is just a dummy article.
Network News Transfer Protocol

- NNTP is used to distribute, inquire, retrieve, and post news articles over a reliable full duplex communication channel (TCP based)
- Defined in RFC 3977
- Communication is done in a request-reply manner after establishing a connection to the server
- Client sends text-based commands and optional parameters to the server and server answers with a result consisting of a status code (e.g. successful execution) and a command dependent part
- Examples for commands are (COMMAND parameter):

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROUP somegroup</td>
<td>Selects newsgroup and returns summary information about it to the client</td>
</tr>
<tr>
<td>POST</td>
<td>Initialises the posting of an article from a client to a news server</td>
</tr>
<tr>
<td>ARTICLE [msg-id</td>
<td>article number]</td>
</tr>
<tr>
<td>IHAVE message-id</td>
<td>Indication for server that another server has an article with given id</td>
</tr>
</tbody>
</table>
Usenet Server – Server communication

Transit communication (flooding algorithm)

1. Message is posted to a newsgroup on one news server

2. News server deploys a data structure that is used to look up all neighbouring servers that host the associated newsgroup

3. Message-ID is sent to relevant servers with the use of the “IHAVE” command; receiving host checks if it does not have this article already

4. If server does not have it, a status code is returned to indicate this fact and the article is sent to the receiving host

5. Procedure is repeated
Usenet Client – Server communication

Two important command types:

1. Submitting articles

   - Establishing a connection to a news server
   - User sends an article to a newsgroup that is available on this server

2. Reading operations

   - After establishing a connection the client requests for different information in defined intervals or on user activity as e.g. for:
     - Available newsgroups
     - Summary of articles
     - One concrete article (by id)

   - If requested information is available, it is transferred to client
Reading operations

Status code
(initial "2" == success)

Lists all available newsgroups

“news.announce” has articles with article numbers from “3000234” to “3002322” and posting is permitted (“y”)

Chooses one group

Chooses the current article

Message header

Message content

[C] LIST
[S] 215 list of newsgroups follows
[S] comp.risks 441970 441099 n
[S] alt.rfc-writers.recovery 4 1 y
[S] tx.natives.recovery 89 56 y
[S] news.announce 3002322 3000234 y
[S] .

[C] GROUP news.announce
[S] 211 1234 3000234 3002322 news.announce

[C] ARTICLE
[S] 220 3000234 <45223423@example.com>
[S] Path: sysx!serv5!sys3!mars!harry
[S] From: From: harry@mars.dom.COM (Harry Hopper)
[S] Newsgroups: news.announce
[S] Subject: Important news notification
[S] Date: Fri, 17 Nov 1970 17:05:42 GMT
[S] Organization: Mars and Son Inc.
[S] Expires: Thu, 1 Jan 2030 00:00:00 -0500
[S] Message-ID: <45223423@example.com>
[S] 
[S] This is just a dummy article.
[S] .
Submitting articles

1. Server does not allow posting for a specific client or for a group of clients:

   [S] 201 NNTP Service Ready, posting prohibited
   [C] POST
   [S] 440 Posting not permitted

2. Server permits posting for a specific client or for a group of clients:

   [S] 200 NNTP Service Ready, posting permitted
   [C] POST
   [S] 340 Input article; end with <CR-LF>.<CR-LF>
   [C] From: "Harry Hopper" <harry@mars.dom.COM>
   [C] Newsgroups: news.announce
   [C] Subject: Important news notification
   [C] Organisation: Mars and Son Inc.
   [C]
   [C] This is just a dummy article.
   [C] .
   [S] 240 Article received OK

• After receiving the article the server has to assign an unique id to it
Web based subscription services

- (X)HTML is not applicable to represent information in an uniform way making data extractable by client software

```
2 of British Team Conquer Everest

May 29, 1953
Katmandu, Nepal
The British expedition has conquered Mount Everest, a radio message flashed from Namche Bazar to the British Embassy here said today. The message said Edmond Hillary, a New Zealand beekeeper and mountaineer, and Tensing Norkay, the famous Sherpa guide, had reached the hitherto unscaled summit from Camp Eight last Friday.

July, 21st, 1969
Men Walk On Moon
Houston, USA
Two Americans, astronauts of Apollo 11, steered their fragile four-legged lunar module safely and smoothly to the historic landing yesterday at 4:17:40 P.M., Eastern daylight time.
```

Additional format is necessary
Web feeds

- A web feed is a special data format used to make web content available for subscription
- Normally provided in additional file that is referred by web page and that contains references to web content as well
- Accessible through HTTP
- Users subscribe to web feeds by the use of feed reader software (aggregator)
- Uniform representation of content makes bundling of information from different web sites possible (syndication of content)
- Most important feed formats are RSS and Atom
RSS

- RSS refers to a set of incompatible versions of XML based Web content syndication formats
- Some of them are:
  - Resource Description Framework Site Summary (RSS 0.9 and 1.0)
  - Rich Site Summary (RSS 0.91, RSS 1.0)
  - Really Simple Syndication (RSS 2.0)
- In the following RSS stands for Really Simple Syndication
- Feeds are organised into channels which may contain any number of items
- Every item has to contain a title or a description
- An item may contain a link to an associated web resource
- Item-level content may be:
  - escaped HTML (e.g. “&lt;/a&gt;” instead of “</a>”)
  - plain text, with no markup
- Often used MIME-type: application/rss+xml
RSS example

```xml
<?xml version="1.0" ?>
<rss version="2.0">
  <channel>
    <title>example.org web feed</title>
    <link>http://example.org/webpage</link>
    <description>Web feed of the example.org site.</description>
    <pubDate>Tue, 12 Jun 1970 04:00:00 GMT</pubDate>
    <item>
      <title>New software version available.</title>
      <link>http://example.org/webpage/1970/06/19/1.html</link>
      <description>After a long time of development version 1.0 of our software has been released.</description>
      <pubDate>Tue, 19 Jun 1970 09:39:21 GMT</pubDate>
    </item>
    <!-- More items...
    <item>...
    </item> -->
  </channel>
</rss>
```
Motivated by the existence of too many incompatible versions of RSS and their lack of standardisation, the alternative XML based syndication format Atom was developed.

Described in RFC 4287

Atom feeds are organised in a number of entries.

Every entry has to contain an id, a title and the time of update.

Text constructs inside tags may be:
- plain text, with no markup (default, type="text")
- escaped HTML (type="html")
- well-formed XHTML markup (type="xhtml")

Beside these the content element might contain pure XML or Base64 encoded text.

Defines special data structures as e.g. a Person construct to specify the author of the feed by giving his name, e-mail address etc.

MIME-type: Content type application/atom+xml
Atom example

```xml
<?xml version="1.0" encoding="utf-8"?>
<feed xmlns="http://www.w3.org/2005/Atom">
  <title>example.org web feed</title>
  <author><name>Harry Hopper</name></author>
  <id>http://example.org/webpage</id>
  <updated>1970-06-19T04:00:00Z</updated>
  <entry>
    <title>New software version available.</title>
    <link href="http://example.org/webpage/1970/06/19/1.html"/>
    <id>http://example.org/webpage/1970/06/19/1.html</id>
    <summary>version 1.0 released.</summary>
    <content type="xhtml">
      <div xmlns="http://www.w3.org/1999/xhtml">
        After a long time of development version 1.0 of
        our software has been released.
      </div>
    </content>
  </entry>
  ... </entry>
</feed>
```
Feed indication

- Availability of web feed can be indicated to user by a simple link or by displaying an special icon
- Atom supports a standardized way of autodiscovery of a (X)HTML link to a web feed; RSS autodiscovery is based on the same principle and is commonly used and supported by many feed readers

```
<html>
  <head>
    <link rel="alternate" type="application/atom+xml"
           href="http://www.example.com/webpage/index.atom"/>
  ...
```
Feed generation

- Out of extracted web content or explicitly stored feed-only content the feed document is created and afterwards optionally validated.
- Feed document can be generated by using:
  - Simple string operations (concatenating tags and content)
  - Regular XML tools (parsers etc.)
  - Special purpose libraries (such as e.g. ROME for Java)
Feed generation

- Beside explicit representation of published information it is possible to generate information from associated (X)HTML documents automatically.

- Information can be based on
  - Definition of strict conventions for the structure of a (X)HTML document thus making relevant information discoverable and extractable by a parser
  - File system information (e.g. modification time, file size)
  - Globally defined information (e.g. name of website owner)
  - Use of HTML meta tags

```html
<html>
<head>
<meta name="summary" content="Some content"/>
</head> ...
<entry>
<summary type="text">Some content</summary>
...</entry>
```
Feed Readers

- RSS and Atom Feed readers are simple XML parsers that have many forms of occurrence as e.g.:
  - Separate software
  - Part of the Web browser
  - Web based readers
  - Integrated into the desktop
- Parsed feed document can be easily organised for further processing, evaluation or displaying in a hash table
Efficiency aspects

- Common techniques to reduce bandwidth consumption or avoid downloading of feeds that have already been downloaded are:
  - Compression
    - Web servers can compress content e.g. by the use of gzip
  - Delta encoding in HTTP (RFC 3229)
    - Mechanism for transferring a description of changes, rather than the entire new instance of the resource; can be used to request only new items of a feed
  - HTTP conditional GET
    - Included into the HTTP request feed reader sends last-modified date of a feed and server only returns new data if no modification was made since this date
Podcasting

- Multimedia file distribution over Internet using web feeds for referencing downloadable files
- Podcasting clients are simple feed readers that support file reference extraction from the feed, download of that file and often specification of player software that is automatically associated with downloaded files (e.g. by adding them to its play list)
Podcasting with RSS 2.0

• File reference is given inside an item by using the element `<enclosure>` that specifies an URL, file length and content type
• Only one `<enclosure>` element per item is allowed
• Beside multimedia files arbitrary file formats can be made available by this indication

```xml
<item>
  <title>Lecture as Video Podcast</title>
  <pubDate>Sat, 11 Nov 1970 00:00:00 +0100</pubDate>
  <description>Lecture number 7 as Podcast</description>
  <enclosure url="http://somedomain.com/files/lecture7.mp4" length="51764473" type="video/mp4" />
  ...
</item>
```
Podcasting with Atom

• File is referenced inside an entry by the <link> element
• More than one file can be referenced thus making it possible to differentiate between file formats, content languages etc.

```xml
<entry>
  <title>Lecture as Video Podcast</title>
  <link rel="enclosure"
       href="http://somedomain.com/files/lecture7_ger.mp4"
       hreflang="de-DE" type="video/mp4" length="51764473"/>
  <link rel="enclosure"
       href="http://somedomain.com/files/lecture7_en.mp4"
       hreflang="en-GB" type="video/mp4" length="50462473"/>
  <updated>1970-12-13T18:30:02Z</updated>
  <summary>Lecture number 7 as Podcast </summary>
  ...
</entry>
```
Further feeds applications

- Web feeds can be used to monitor different information by making them available through a subscription service such as:
  - Summary of new Mails in E-Mail-Inbox
  - Tracking changes in source code version control systems as e.g. Subversion
  - Events that occur in a server log file
  - Usenet messages
Further feeds applications

- Information offered by Web Services can simply be provided for subscription (e.g. for monitoring offers at an auction platform in a feed reader)
- Intermediate server requests for information by using the Web Service interface and converts the result to a web feed format
- Conversion to web feed can be done by Extensible Stylesheet Language Transformation (XSLT)
Conclusion

<table>
<thead>
<tr>
<th>name</th>
<th>server</th>
</tr>
</thead>
<tbody>
<tr>
<td>news1</td>
<td>newsserver.com</td>
</tr>
<tr>
<td>weather1</td>
<td>weatherserv.com</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Distributor aware subscription

User client aware subscription

Feed reader

Feed

RSS/Atom over HTTP

NNTP

Usenet client

Mail client

Server

User information

(newsgroup1
  newsgroup2
  ...
)

Subscribed to

(reference
  (multimedia) file)
References

Links at the IETF


Further links

RSS 2.0 specification http://www.rssboard.org/rss-specification
ROME RSS/Atom library http://rometools.github.io/rome/