HELSINKI UNIVERSITY OF TECHNOLOGY
Faculty of Electronics, Communications and Automation
Department of Media Technology

Chip Gylfe

# METADATA IN CROSS MEDIA EDITORIAL PROCESSES

Master's Thesis

Espoo, July 7th, 2009

Supervisor: Professor Pirkko Oittinen

Instructor: Stina Westman, M.Sc. (Tech)

| | |
|---|---|
| TEKNISKA HÖGSKOLAN<br>Fakulteten för Elektronik, Kommunikation och Automation<br>Utbildningsprogrammet för Telekommunikationsteknik | SAMMANDRAG AV DIPLOMARBETET |

| | | |
|---|---|---|
| Utfört av: Chip Gylfe | | |
| Arbetets namn: Metadata i Cross Media Redaktionsprocesser | | |
| Sidantal: 84+19 | Datum: 7.7.2009 | Språk: Engelska |
| Professur: Medieteknik | Kod: AS-75 | |
| Övervakare: Professor Pirkko Oittinen | | |
| Handledare: Stina Westman, DI | | |


Olika typer av media är inte längre bundna till specifika distributionskanaler i samma grad som de en gång varit. Nya krav på mediainnehåll samt -tillgänglighet har lett till att företag inom mediabranschen måste anpassa sig genom att tillämpa nya rutiner för skapande och administrering av innehåll. Detta inverkar inte endast på arbetsrutiner, utan leder i sin tur till nya krav på DAMS system i vilka metadata kommer inneha en väsentlig roll. Utan välstrukturerad metadata kommer många arbetsuppgifter visa sig extremt tidskrävande.

Den här studien erbjuder en genomförlig och aktuell empirisk forskning med insyn på tillämpningen av metadata inom innehållsproduktionen hos välkända medieföretag som utövar cross media publicering. Företag som deltog i studien kan beskrivas som nyhetsbyråer, tidnings- eller tidsskriftsutgivare.

En abstrakt metadatastruktur bestående av essens, kontextuell, strukturell, och administrativ metadata var konstruerad på basen av uptäckt metadata hos de besökta företagen. Denna struktur avspeglar nuvarande krav på metadata inom cross media företag. Metadatastandarder rekommenderas som bas för kartläggning mellan lämpliga metadatafält och de i strukturen angivna metadata-attributen. På basis av resultaten gavs även förslag på framtida krav på metadata inom cross media redaktionssystem. Metadata var även analyserad på basen av typ, egenskaper, och användning. Subjektiva åsikter om nuvarande metadatavanor togs även i beaktande. Redaktionsprocesser inklusive metadatalivscykeln var modellerade med BPMN.

Bland de företag som deltog i studien uppkom skillnader i processer anknytna till cross media publicering. Alla deltagande parter visade dock ett gemensamt intresse för att förbättra den nuvarande situationen. En annan gemensam överenskommelse uppstod i formen av ett framtida krav på redaktionssystem gällande tema- eller ämnescentrerad planering av innehåll. I synnerhet essensbaserad metadata är i starkt behov av förbättringar för att klara av framtida krav. Medieföretag har insett att investeringar i metadata inte endast minskar på arbetsbördan och är till förmån för administrering av tillgångar, utan också erbjuder totalt nya möjligheter för affärsverksamhet.



Nyckelord: metadata, metadatastruktur, metadatalivscykel, cross media publicering, redaktionsprocess, tidning, nyhetsbyrå, tidsskrift, nyheter, empirisk

HELSINKI UNIVERSITY OF TECHNOLOGY
Faculty of Electronics, Communications and Automation
Degree programme of Telecommunications

ABSTRACT OF THE MASTER'S THESIS

Author: Chip Gylfe

Title: Metadata in Cross Media Editorial Processes

Number of pages: 84+19 | Date: 7.7.2009 | Language: English

Professorship: Media Technology | Code: AS-75

Supervisor: Professor Pirkko Oittinen

Instructor: Stina Westman, M.Sc (Tech)

Different types of media such as text, images, audio, and video are no longer restricted to different publishing channels in the same way that they used to be. Media companies are adopting new routines for handling and creating content to better suit the new demands, which in turn puts new requirements on digital asset management systems (DAMS). Metadata will have an essential role in managing content in these systems, and without well structured metadata many tasks become extremely time consuming.

This study provides an extensive, up-to-date, empirical research of the current use of metadata across the content creation process within well known media companies involved in cross media publishing. Three case companies were selected for this study: a newspaper, a magazine publisher, and a news agency. Observations and surveys, including interviews and questionnaires were conducted at the research sites.

A metadata framework consisting of essence, contextual, structural, and administrative metadata field attributes was created based on the discovered metadata in case companies. The framework serves as a platform for current metadata needs in media companies utilizing cross media publishing. Metadata standards are suggested to be used as a source for mapping suitable metadata fields to the provided metadata attributes. Suggestions for future requirements on metadata in editorial systems were also given based on the results. In addition to the metadata framework, discovered metadata was also assessed based on nature, characteristics, and use. Subjective viewpoints of current metadata practices are also taken into account. Current editorial processes including the metadata time-cycle were modeled with BPMN annotation language.

There were differences in cross media publishing routines between the case companies, but all participants showed a common interest to enhance current practices. A mutual requirement on future editorial systems was the inclusion of theme or topic based planning of content. The results showed that especially essence based metadata needs improvements in order to cope with future requirements. Media companies have realized that investing in metadata not only reduces workload and is beneficial from an asset management perspective, but also provides completely new business opportunities.

Keywords: metadata, framework, life-cycle, cross media publishing, editorial processes, newspaper, news agency, magazine, news, empirical

# ACKNOWLEDGEMENTS

This master's thesis was written at the Department of Media Technology, Helsinki University of Technology. The study was founded by TIVIT and TEKES as part of the CrossMedia Project, which is connected to the Flexible Services Programme 2008.

I would especially like to thank my supervisor, Professor Pirkko Oittinen, for giving me the opportunity to write my thesis as part of the CrossMedia Project. My instructor Stina Westman was always willing to help, and I am grateful for her time.

I would also like to thank TIVIT and TEKES for the funding that they provided, but also all the people from the case companies that participated in this study. Thanks to all colleagues at the department for creating a great atmosphere to work in.

Last but definitely not least I thank my family and my fiancée Sofia Henriksson for their never ending support.

Espoo, July 7th, 2009

Chip Gylfe

# TABLE OF CONTENTS

**1 INTRODUCTION ..... 1**

1.1 Motivation ..... 1

1.2 Research questions ..... 3

1.3 Scope of this thesis ..... 3

1.4 Thesis structure ..... 4

**2 CROSS MEDIA EDITORIAL PROCESSES ..... 6**

2.1 Change from traditional editorial processes towards cross media publishing ..... 6

2.2 Relationship between content creation and metadata time-cycle ..... 6

2.3 Cross media editorial processes ..... 9

2.4 Publishing channels ..... 12

**3 METADATA IN CROSS MEDIA PUBLISHING ..... 13**

3.1 Considerations and scope ..... 13

3.2 Definition ..... 13

3.3 Nature and characteristics ..... 14

3.3.1 Granularity ..... 14

3.3.2 Free-form or controlled values ..... 15

3.4 Metadata types ..... 16

3.5 Standards for multimedia metadata ..... 19

3.5.1 IPTC: NewsML-G2 and NewsCodes ..... 19

3.5.2 Adobe XMP ..... 20

3.5.3 Dublin Core ..... 20

3.5.4 IPTC: Photo Metadata ..... 21

3.5.5 DIG35 ..... 21

3.5.6 EXIF ..... 21

3.6 Relationship between metadata and content management ..... 22

**4 METHODOLOGY ..... 24**

4.1 Methodology for studying metadata in editorial systems ..... 24

4.1.1 Selection of methods and data ..... 24

4.1.2 Methodology for studying legislative documents: RASKE project ..... 24

4.2 Research sites ..... 26

4.2.1 Helsingin Sanomat ..... 26

4.2.2 Sanoma Magazines Finland ..... 27

4.2.3 STT ..... 27

4.3 Analysis – Methods and data ..... 27

4.3.1 Overview ..... 27

4.3.2 Observations ..... 29

4.3.3 Modeling ..... 31

4.3.4 Metadata framework ..... 32

4.3.5 Interviews ..... 36

4.3.6 Questionnaire ..... 37

**5 RESULTS ..... 41**

5.1 Modeling of current editorial processes and metadata life-cycle ..... 41

5.1.1 Planning ..... 41

5.1.2 Production ..... 47

5.1.3 Publishing and distribution ..... 53

5.1.4 Archiving ..... 54

5.1.5 Metadata life-cycle ..... 55

5.2 Metadata framework for cross media publishing ..... 56

5.2.1 Essence based field attributes ..... 57

5.2.2 Structural field attributes ..... 57
5.2.3 Administrative field attributes ..... 57
5.2.4 Contextual field attributes ..... 58
5.3 Characteristics of metadata fields ..... 58
5.3.1 Metadata fields in editorial systems versus archived metadata ..... 58
5.3.2 Value insertion – automation and structure ..... 59
5.3.3 Life-cycle of metadata ..... 62
5.4 Use of metadata fields ..... 62
5.4.1 Fill rate of metadata fields ..... 62
5.4.2 Unique values in metadata fields ..... 63
5.4.3 Comparison of objective and subjective metadata use ..... 65
5.5 Survey on metadata ..... 66
5.5.1 Use frequency of metadata based on content type (Q1-Q5) ..... 67
5.5.2 Creation and modification of metadata (Q6-Q13) ..... 68
5.5.3 Benefits and usefulness of metadata (Q14-Q22) ..... 69
5.5.4 Difficulties related to current metadata (Q23-32) ..... 71
**6 DISCUSSION ..... 73**
6.1 Goals of the study ..... 73
6.1.1 Challenges related to content management and metadata ..... 73
6.1.2 Suggestions for requirements on metadata in future cross media editorial systems 73
6.1.3 Editorial processes and the metadata time-cycle ..... 76
6.2 Reliability analysis ..... 77
6.2.1 Data ..... 77
6.2.2 Results ..... 78
**7 CONCLUSIONS ..... 80**

# 1 INTRODUCTION

## 1.1 Motivation

News and media companies have come a long way in their digitalization process. Today, it is not unusual for media to be created and preserved in digital from until it reaches the end-user and ultimately the archive. Different types of media such as text, images, audio, and video are no longer restricted to different publishing channels in the same way that they used to be. At the receiver end, this increase in publishing possibilities is most visible in the way we consume this information.

As an example, in the past we read the news in the newspaper, listened to it on the radio, and watched it on television. Those were our most common options, and each one of them was usually provided by different companies. Today we not only have multiple ways of receiving news in various formats but current multimedia devices enable cross usage of various publishing channels on the same device (Bødker & Petersen 2007, p. 539-540). The smart phone is a good example of this. We can use it to read the news on the internet, or receive a shorter version of the news as a text message. We could also use the smart phone to listen to the news on the radio, or we could use it to watch the news as a video stream. It should be noted that although we are able to take in this information via different channels, and its visual appearance might differ, the essence often remains the same. However, distribution of the same essence is not a criterion for cross media publishing. This way of publishing content through a variety of channels, be it either content with the same essence or content complementing each other, has given birth to a new term that will be frequently used in this study, cross media.

*"Cross media can be defined as any content (news, music, text, and images) published in multiple media (for example print, Web, and TV)."* (Veglis 2008b)

The above described scenario is possible because of the digitalization and convergence that has occurred, but the definition is not optimal since it lacks to mention the possibility to use different media for complementary purposes. No doubt this has been a challenging time for media companies, and still is, as technology continues to develop at a fast rate. Not so long ago the publishing process was linear, whereas it today is split into multiple segments somewhere in the creation process (Möller 2001). It is important to note, however, that ideally the content and its presentation should be kept separate as long as possible. The further this is possible, the less work is needed to produce different media products based on the same content (Jokela 2001, p. 53).

Media companies have realized the need for renewal of processes and reacted by changing their infrastructure and by adopting new routines for handling and creating content to better suit the new demands. This is no small task, and partly due to the major challenges of these adjustments, some aspects have not been planned thoroughly. Since most media content is stored in digital form and the pace at which new content is created is increasing day by day, this introduces completely new challenges on archiving the content. Or more

correctly, challenges on making the retrieval of relevant content possible. The solution comes in the form of metadata:

*"The purpose of metadata is 'to facilitate search, evaluation, acquisition, and use' of resources"* (IEEE 2001; according to Duval 2001).

At the moment it seems that one of the main challenges within media companies is related to handling their information assets. Descriptive data about data, i.e. metadata, is a topic that has received a lot of attention. Implementing metadata efficiently is, however, not an easy task, and current metadata within organizations might be insufficient because of, for example, poor structure or missing data elements. Older data in the archives might be missing metadata altogether. There is a clear need for organizing, structuring, and describing content in such a manner that we can retrieve relevant data with minimal effort. If done properly this wealth of information could easily be accessed with appropriate processes. Without suitable metadata the retrieval of relevant data is very hard and time consuming, which in turn generates additional expenses. This is especially true for media in the form of images, video and audio.

Descriptive data is not only important for the retrieval of content. If used correctly it supports the creation process and every participant involved, and even the end-user can benefit from the metadata. Rights and protection information are also becoming more and more important when managing multimedia data. This type of information could be embedded or linked to the content (Pereira et al. 2008, p.731). This should take place within the content creation process. Another key benefit of using metadata within media companies is related to optimizing cross media publishing and minimizing manual input during the process.

Simply agreeing on attaching or linking metadata to the content is not enough. Careful consideration about metadata types, semantics, and the life-cycle of metadata is required, because a poorly chosen metadata structure will only lead to more difficulties than benefits. Available standards should also be considered, although they might not always suffice for specific needs.

The value of descriptive data has been overlooked partly due to other challenges and costs that have arisen because of the rapid transfer from analog to digital and the ever increasing pace of new media generation. The amount of data that the media companies currently have hold of is immense, and it grows constantly. Chakravarty et al. state in their research that some large organization's intranets have reached the size of mini webs, and that their size soon might be comparable to the Internet at the end of the 90s (2004, p. 1).

This study provides an extensive, up-to-date, empirical research of the current use of metadata across the content creation process within well known media companies in Finland. The involved media companies are a newspaper, a news agency, and a magazine publisher.

This study is part of the Cross Media Publishing project which is funded by the Finnish Funding Agency for Technology and Innovation (TEKES) and Tivit. The goals of the project that are related to this study include achieving the following:

- extensible metadata systems within next generation editorial systems, which unite professional and user needs
- business models and demonstrator system for a flexible supply chain allowing capacity and resource management in production and distribution of customized media products and services
- cross media solutions for life-long learning that enable the tailoring of the personal learning environment, while increasing the use of hybrid media

The executive summary of the project can be found in Appendix 1.

## 1.2 Research questions

The research questions focus on resolving challenges related to metadata within editorial operations in media companies. In order to answer the research questions, an extensive inventory of currently used types of metadata is needed, as well as a time-line representation of the creation and modification of metadata. When these initial stages are completed, focus can be shifted towards areas for improvements. The technical implementation of suggested improvements is out of the scope of this study.

1) What *challenges* do media companies face today with regard to *content management* and *metadata*?

2) *What types* of metadata elements are currently used, and *do they form a logical structure*?

3) *How*, *by who*, and *when*, is metadata *created* or *modified* within the content creation process in media companies?

4) How could the metadata structure and the creation of metadata within the editorial processes be improved?

## 1.3 Scope of this thesis

This thesis provides suggestions for improvements on metadata implementations in current editorial systems, but does not focus on the technical realization of them, and neither does it include field experiments or simulations. Future tasks in the CrossMedia project will however utilize these results when designing improvements for current editorial systems. By studying Nunamaker et al.'s (1991; according to Lyytikäinen V. 2004, p. 26) multimethodological research approach, resemblances between their model and different parts of the CrossMedia project can be identified. This thesis can be seen as part of the *theory building* and *observation* activities of their model.

*"Theory building includes development of new ideas, concepts, frameworks, methods, or models."*

*"Observation includes case studies, field studies and sample surveys that are unobtrusive research operations."*

Later research in the CrossMedia project will fall into the categories of *experimentation* and *systems development*. Nunamaker et al.'s model may be seen as a very low-level, abstract visualization of how this thesis is connected to the overall work of the CrossMedia Publishing Project. An illustration of Nunamaker et al.'s model is presented in Figure 1. It should be noted that there are also other elements in the project that fall into the same categories, but these will not be discussed here.

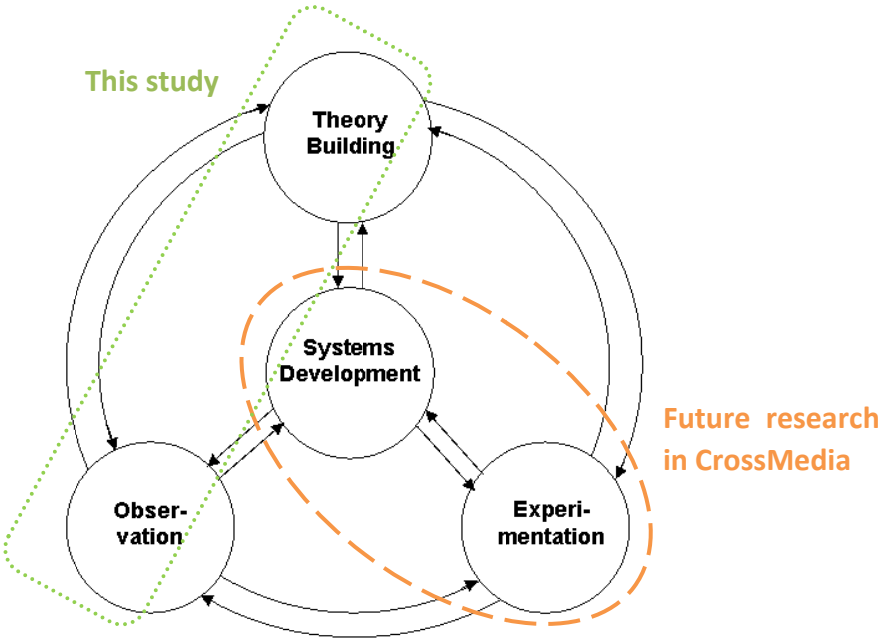



**Figure 1. Multimethodological research approach (Nunamaker et al. 1991, p. 209; according to Lyytikäinen V. 2004, p. 26)**

The observation in this study includes case studies, interviews, and questionnaires. As a result, models of the current implementation of metadata in editorial processes are produced, including suggestions for improvements based on these models.

## 1.4 Thesis structure

Related literature is discussed in Sections 2 and 3. The first part of the literature review gives an overview of cross media editorial processes. The relationship between metadata and content creation is also discussed. Metadata is in main focus in the literature review, including different views on metadata definition, nature and characteristics of metadata, and reasons for media companies to invest in metadata. Types of metadata as well as standards suitable for multimedia metadata are also considered.

Section 4 presents the methodology used in this study. The selection of suitable methods is discussed, as well as the phases of the analysis. Qualitative and quantitative data used in this study are presented, including documents and data collected through observations, interviews and questionnaires. Approaches used for analyzing the data are discussed. Research sites are also presented in this section.

Section 5 contains the results of the study. Models of current editorial processes and metadata life-cycle are first presented, followed by a metadata framework which serves as an abstraction of current metadata needs in media companies. Different types of discovered metadata types and overall use of metadata fields are also discussed. Finally subjective views on current metadata practices are presented.

Discussion in Section 6 examines if the goals of the study were reached. Reliability of the data and the results are assessed. Suggestions for requirements on metadata in future editorial systems are also given. The final section of this thesis presents the conclusions of the study.

# 2 CROSS MEDIA EDITORIAL PROCESSES

## 2.1 Change from traditional editorial processes towards cross media publishing

The focus of this study is on metadata in cross media editorial systems, and the basic principles of cross media publishing therefore need to be discussed. The term *editorial office* will frequently be used in this study. It refers to the place in the media company where the media content is created, handled, or where it is submitted before being published. The terms *editorial system* and *editorial software* will also be used. They both refer to software that is used to control, handle, and create content in the editorial office. The term *editorial process* covers the editorial system, work routines and practices, and the actors involved in creating and ultimately publishing the content.

The main difference between traditional publishing and cross media publishing is the utilization of various publishing channels. However, the word *cross* does not necessarily imply that the same content needs to be distributed across multiple channels. Content is tailored to suit the format of the publishing channel and different content can be distributed via separate publishing channels in order to complement each other. This raises new requirements on the editorial processes including the involved metadata. This section discusses possible publishing channels, involved editorial processes, and the actors within these processes.

## 2.2 Relationship between content creation and metadata time-cycle

Jokela (2001, p. 53-57) stated that in an ideal situation the content essence and its presentation should be kept separate as long as possible. The further this is possible, the less work is needed to produce different media products based on the same content essence. It is important to note at this point that content is not the same as the essence of content. The Society of Motion Pictures and Television Engineers (SMPTE) (according to De Stutter et al. 2006 p. 221) states that:

$$content = essence + metadata$$

In other words, content does not exist without essence and metadata. Metadata is discussed in more detail in Section 3.

Jokela (2001) states that in order to achieve greater reusability, required characteristics should be formalized into reusable media component- and product templates. In addition, Jokela highlights the importance of implementing user input in the planning phase of content creation. This is also seen as an important part of future editorial systems by the CrossMedia Publishing Project. Figure 2 illustrates these characteristics.

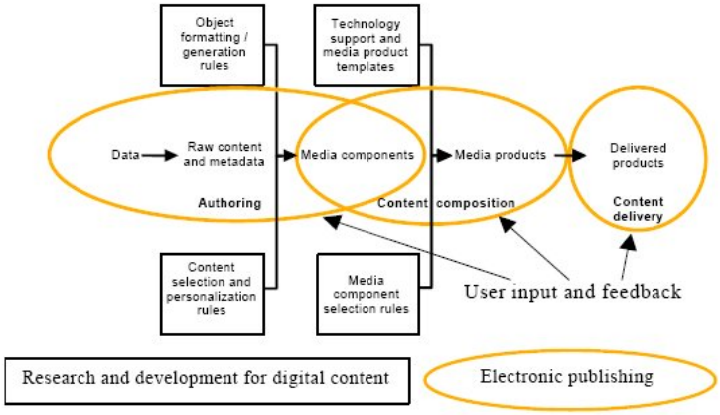



**Figure 2. Key processes and process steps for publishing digital content (Jokela 2001)**

Additional means are needed in order to efficiently produce content in the above stated manner. A traditional linear push workflow model for broadcasting is presented by Mauthe & Thomas (2004) in Figure 3. Although it represents a workflow model for broadcasting the idea is applicable for other content creation and publishing processes in current media companies. The problem with the model is however that it is entirely sequential, which makes content and information about the production process only reachable from each specific work phase.


**Figure 3. Linear push workflow model (Mauthe & Thomas 2004)**

By reorganizing the content management system (CMS) into supporting a content-centric pull workflow design, the CMS becomes a central hub for all content related information that is created during the process. This approach makes the information accessible at any time during the content creation process. This workflow model is presented in Figure 4 (Mauthe & Thomas 2004).

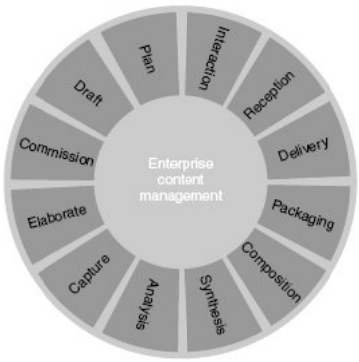



**Figure 4. Content-centric pull workflow model (Mauthe & Thomas 2004)**

In order to truly make the content accessible at any phase in the workflow, and also support the features mentioned earlier by Jokela (2001), an additional flow of metadata is essential. In order for CMSs to support advanced features, metadata should evolve during the content creation processes. Curtis et al. (1999) stated that validation of data and metadata should occur throughout the content lifecycle. Otherwise valuable information is lost or fragmented, and if retrieved later on the initial meaning or purpose of it might have changed. Chakravarty et al. noted that if metadata is filled in based on someone's memory, it might be false or represent a "modified view" of what it actually should be (2004, p. 2). Not only may valuable information be lost forever, but it is also expensive and time consuming to attach metadata at a later stage in the production process (Besser & Malssen). Mauthe & Thomas' (2004) model of essence and metadata flow is presented in Figure 5, which clearly emphasize the importance of a dynamic metadata flow. Mauthe & Thomas' model should not be seen as a de facto model for applying metadata to cross media publishing. It is presented in order to raise thoughts about the benefits and impact that implementing a dynamic metadata workflow may have on the value of assets.

Mauthe & Thomas (2004) are not the only ones pointing out the importance of continuous metadata flows. Chakravarty et al. (2004, p. 2) note that annotation should happen every time a document is accessed, as the content may serve multiple purposes based on user type or needs. Chakravarty gives an example: the author may want to make the document available via ontology-based annotation, and the author generally has a specific view on the reasons why a document is produced and successively retrieved. However, the reader may have very different annotation needs. Furthermore, different users may want to comment on the document itself, or on other comments.

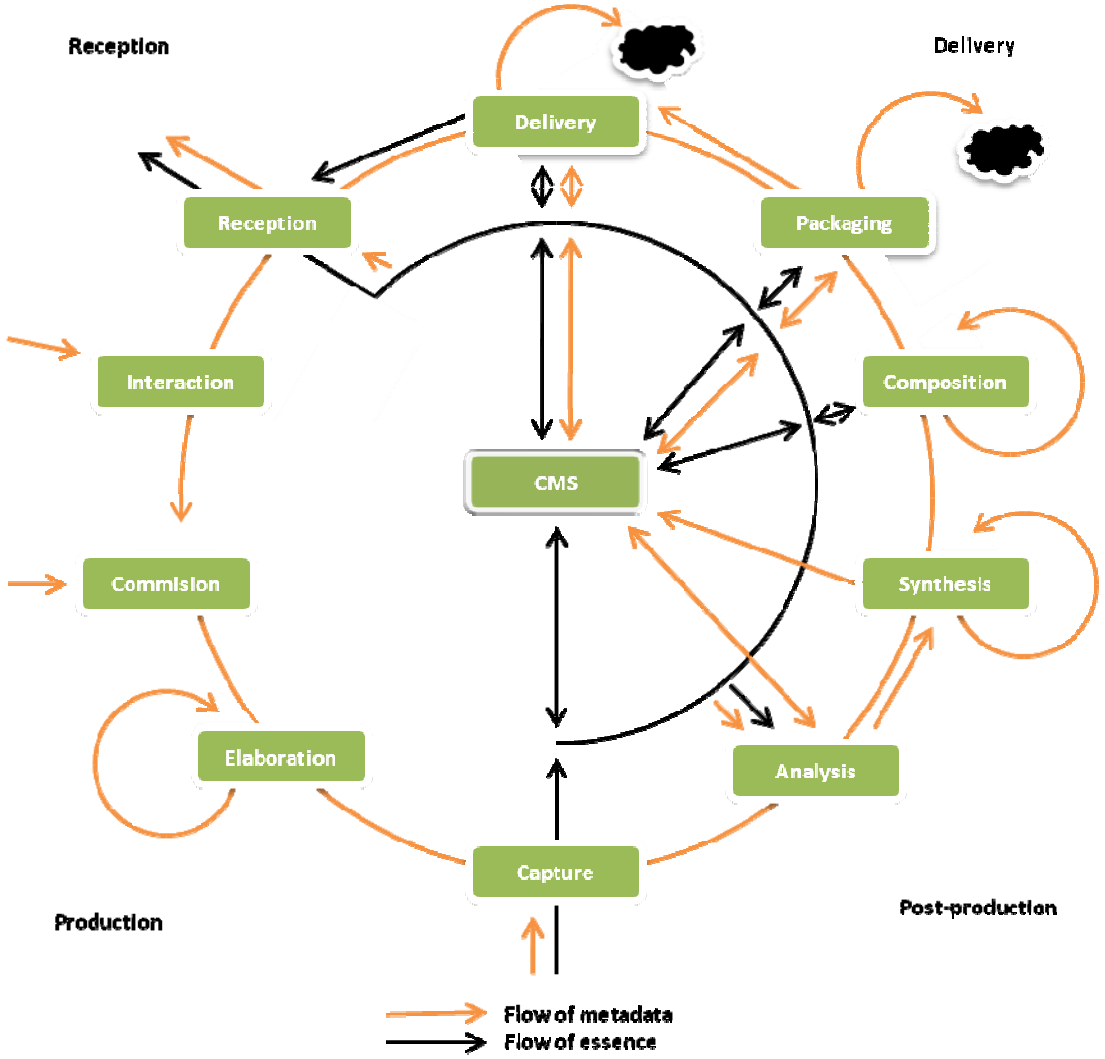



**Figure 5. Essence and metadata flow in a content-centric process model (Mauthe & Thomas 2004)**

## 2.3 Cross media editorial processes

The focus of cross media publishing is in the utilization of various publishing channels. This could mean that the same content is tailored for different publishing channels, but the term is certainly not restricted to this type of practice in this study. Möller (2001) presents a generic workflow model for cross media publishing, which is presented in Figure 6 below. This workflow model is relevant for this study, but it is not optimal as it lacks the possibility to utilize cross media publishing for content complementing each other. Due to this restriction the model is also insufficient in terms of planning phases, as cross media publishing needs to be taken into account from the very beginning of the editorial process.

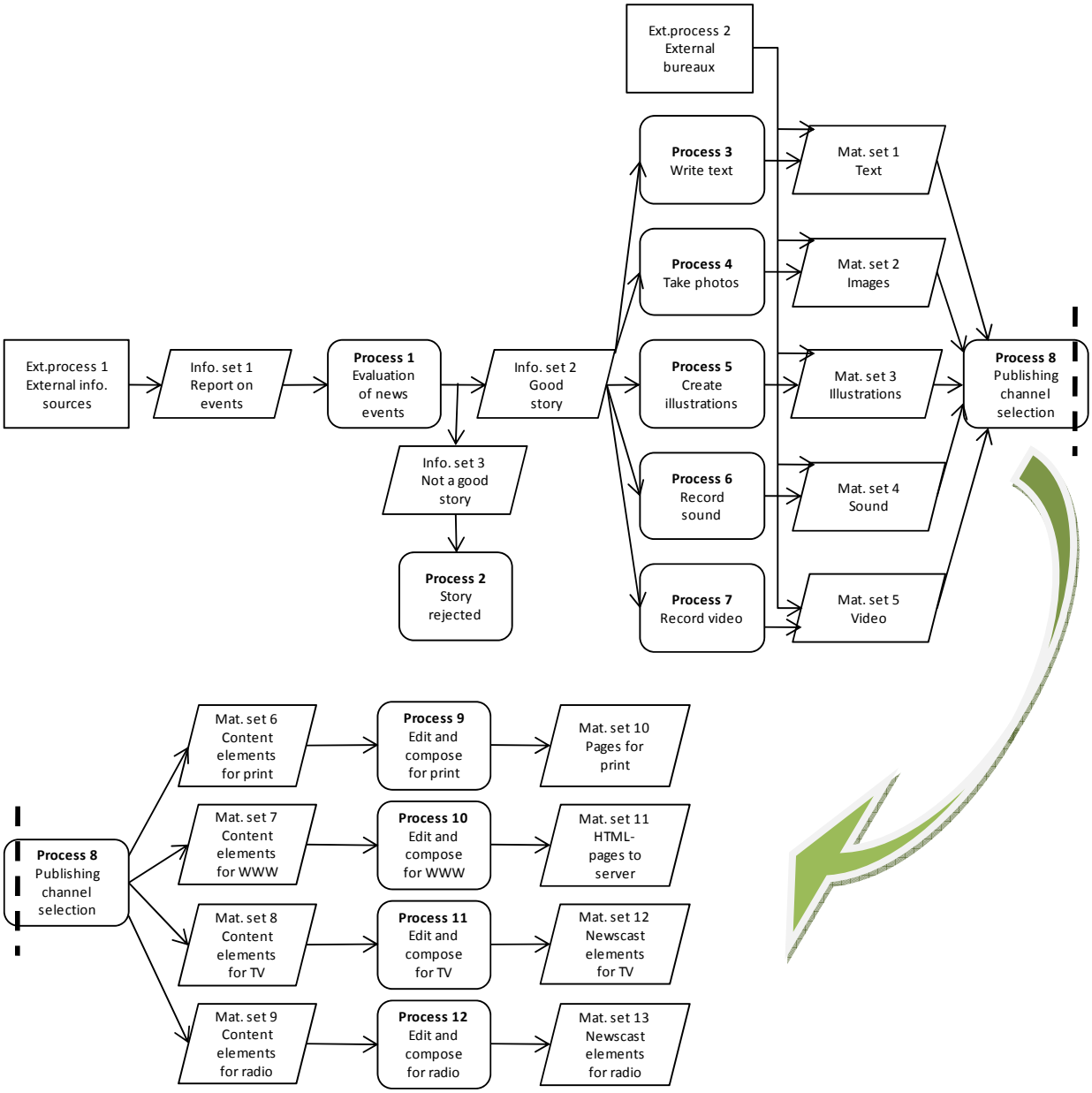



**Figure 6. Generic workflow model for multiple channel publishing (Möller 2001, p. 30)**

Möller (2001) notes later that evaluation of news information in the beginning of the editorial process should include a choice of primary publishing channel. This way of thinking has greater end value potential, as all material cannot be equally suitable for all publishing channels. It should however not restrict content to a single channel. The model shown in Figure 7 does not take into account means for using various publishing channels in order to complement different content, but it does illustrate the current timeline for cross media publishing in a better way than the model in Figure 6. As can be seen in Figure 7, some publishing channels better suit specific media than others. The maturity of the content is also reflected in the choice of publication channel. These factors among other things generate the typical timeline for different publishing channels, which is visualized in Figure 7. Differences between publishing channels are discussed in more detail in the next section.

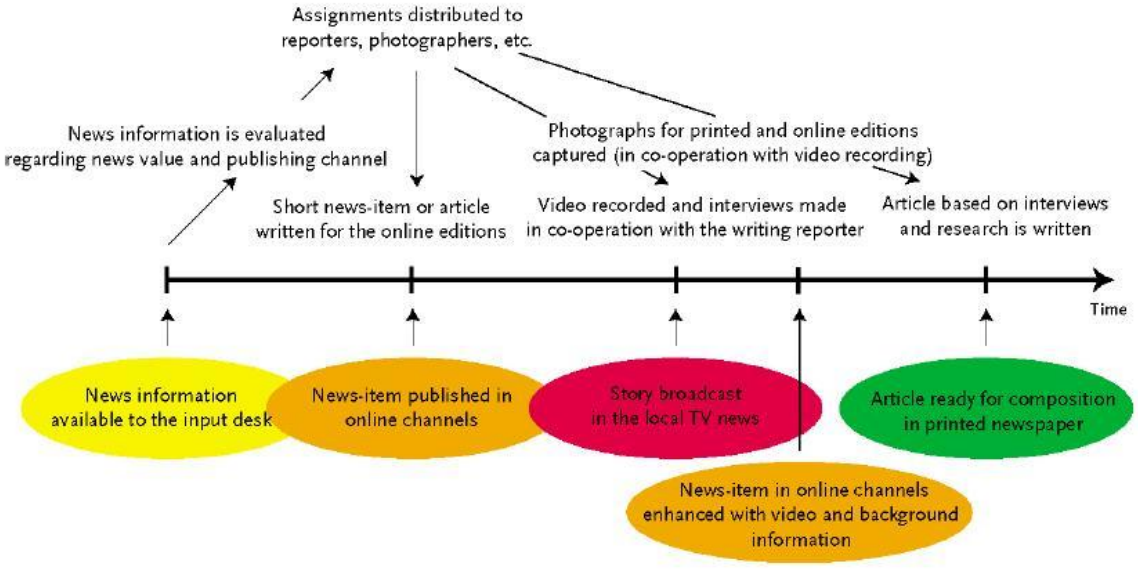



**Figure 7. Example of workflow for content creating within a multiple channel news organization (Möller 2001, p. 31)**

Dailey et al. (2003) present a conceptual framework for understanding convergence, which they call the Convergence Continuum (Figure 8). This model is more abstract from a workflow perspective and does not as such show specific phases of the content creation, but it does show how cross media publishing can be used for complementary purposes. *Cross promotion* is a means for influencing viewers or readers to use another medium for complementary or additional content. *Cloning* defines content that is published in multiple channels without alteration. *Coopetition* occurs when partners compete for creating unique content, but cooperate to some degree so that both partners benefit from the cooperation. For example, both partners could be covering the same story, but the end product is published via different channels. *Content sharing* occurs when partners regularly collaborate by exchanging ideas or forming joint projects. Finally, *convergence* represents the level at which partners have a shared assignment or editor's desk. The partners then proceed by creating content independently in a way that is best suited for a specific publishing channel.

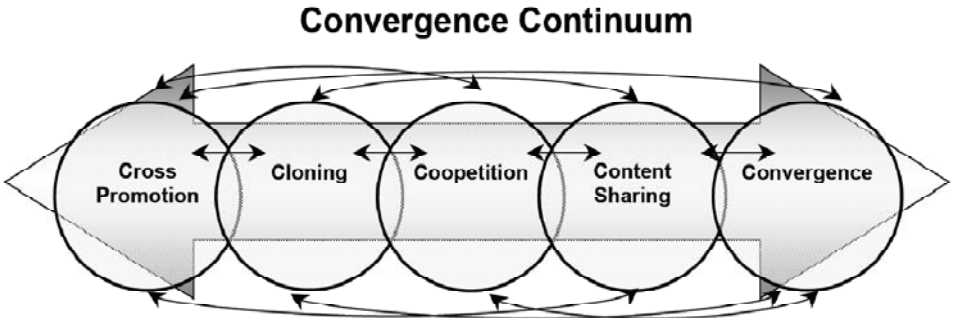



**Figure 8. The Convergence Continuum shows a range of behaviors and their interaction during content production, and illustrates various ways to increase the potential of media assets via cross media publishing (Dailey et al. 2003)**

## 2.4 Publishing channels

Publishing channels represent distinct networks for publishing media content, not necessarily different categories of technology. Some channels might overlap each other, but they are still commonly referred to as separate channels. A more in depth description and comparison between such channels can be found in the Veglis' papers (2008a; 2008b). An overall view of publishing channels used by newspapers around the world is presented in Table 1.

**Table 1. Medium, push-pull orientation, and dynamics of channels (Veglis 2008b)**

| Channel | Medium | | | | Push | Pull | Content | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Static | | Dynamic | |
| | Paper | Storage | Internet | Mobile | | | Text | Picture | Sound | Video |
| **Print** | x | | | | | x | x | x | | |
| **CD/DVD-ROM** | | x | | | | x | x | x | x | x |
| **WWW** | | | x | | | x | x | x | x | x |
| **E-mail** | | | x | | x | | x | x | | |
| **RSS** | | | x | | x | | x | | | |
| **Wi-Fi** | | | x | | x | x | x | x | x | x |
| **Webcasting** | | | x | | | x | x | x | x | x |
| **Blogs** | | | x | | | x | x | x | x | x |
| **PDA** | | | x | | x | x | x | x | x | x** |
| **TabletPC** | | | x | | x | x | x | x | x | x |
| **SMS** | | | | x | x | | x | | | |
| **WAP** | | | x* | x | x | x | x | x | | |
| **PDF** | | x | x | | x | x | x | x | | |

* although WAP is a channel that is implemented on a mobile network the actual service is Internet oriented.
** probably low quality video

From the above table it is instantly clear that cross media publishing has the potential to offer a great deal of content differentiation to the receivers. Some channels are more mobile than others. Channels also differ in how the content reaches the user. Content is sent to the users in push based channels while pull based channels rely on activity from the users side. Channels may support dynamic, static, or both types of content. Publishing channels should not merely be chosen based on these properties, as channels vary greatly in the number of active users. Veglis (2008b) also mentions that the preparation of dynamic content always is more time consuming than creation of static content. This in turn raises questions regarding time and speed requirements. RSS and SMS channels are commonly characterized as info-alerts, whose purpose is to notify the receiver to seek alternative publishing channels for more extensive information (Veglis 2008b). When publishing channels are used to distribute different versions of the same content, the content is usually enhanced in between publications (Veglis 2008b). When this type of momentum based channel separation is performed, the editorial system should support this thoroughly in order to reduce the needed workload amount. The importance of this tightly coupled synergy between the metadata time-cycle and content creation is discussed in more detail in Section 2.2.

# 3 METADATA IN CROSS MEDIA PUBLISHING

## 3.1 Considerations and scope

A wide range of issues related to metadata, such as definition, categorization, characteristics, and standards, have all been discussed in previous literature to a great extent. For the purpose of this study, related to current use of metadata within media companies, a review of some of the basic properties that metadata can offer is needed. This will establish a foundation for the following sections of this thesis.

It should be noted that the main focus is on metadata aspects that directly relate to the implementation of metadata within media companies, such as newspapers, news agencies, and magazine publishers. As an example, there exist a large number of metadata standards, but those presented in this thesis are considered suitable for media companies (see Section 3.5 for more details). Section 2 should therefore not be viewed as a comprehensive literature review or overview of metadata, but as a summary of metadata related matters that are important for cross media publishing and other editorial processes. It should also be noted that this section does not cover technical aspects of how metadata is implemented.

## 3.2 Definition

What is metadata, and how is it defined? One frequently used, loose definition is "data about data". Table 2 serves as an example for the differences in opinion on the definition of the concept.

**Table 2. Different definitions of metadata.**

| Definition | Source |
|---|---|
| *"the sum total of what one can say about any information object at any level of aggregation"* | (Gilliland-Swetland 1998, p. 1) |
| *"information about an electronic document, resource, or the operation of a computer system"* | (Berghel 1999) |
| *"information associated with a piece of data"* | (Curtis et al. 1999, p. 2) |
| *"information about content essence"* | (Jokela 2001, p. 19) |
| *"information about an object, be it physical or digital"* | (IEEE 2001; according to Duval 2001) |
| *"metadata are basically descriptive data"* | (Duval 2001, p. 591) |
| *"optional structured descriptions that are publicly available to explicitly assist in locating objects"* | (Dimitrova 2004, p. 13) |
| *"a system's data dictionary, capturing definitions of data entities and the relationships among them"* | (Shankaranarayanan & Even 2006, p. 90) |

The descriptions vary to a great extent. Some of them imply that metadata only is a means for finding specific data; everything else is data, not metadata. Others describe metadata as

information about an object, i.e. as descriptive data. There even exist differences of opinion whether metadata can be associated with physical or digital objects. Based on the quotes above it is clear that there is a high amount of dispute about what metadata actually is. There is no unique, commonly accepted description of it. Dimitrova (2004, p. 13) comments on this uncertainty as follows: *"for something purported to be essential to capturing all of human understanding, this is not a good start"*.

Perhaps the lack of a common definition simply is due to the complexity of its meaning. It is hard to describe something which by itself is exceptionally descriptive. Shankaranarayanan and Even (2006) comment on this (see their quote above) as follows: "*while not inaccurate, this is also an overly narrow view that overlooks the richness and complexity of metadata"*. Instead of trying to describe metadata, one could ask what its *purpose* is:

> *"to facilitate search, evaluation, acquisition, and use of resources"*
> (IEEE 2001; according to Duval 2001)

This description is very hard to argue against. It covers many of the above mentioned descriptions, or more correctly, their purpose and meaning, in a very compact sentence. It combines a lot of different views, but still manages to be precise. This sentence captures the essence of metadata. *In this thesis metadata refers to essence based, administrative, structural, and contextual information, which is attached or linked to content with the aim of facilitating search, evaluation, acquisition, and use of resources*.

One other matter that needs to be discussed is the relationship between metadata, essence, and content. It is widely accepted that content consists of essence and metadata. Essence refers to the actual raw material, while metadata is used to describe the essence (Mauthe & Thomas 2004). The idea behind this division is that content cannot be effectively processed in content management without knowledge of content essence, which is gained via metadata (Mauthe & Thomas 2004). This concept is used as a building block in this study, but the division is not completely suitable for content in current editorial systems in media companies. Metadata describing content essence is only part of the needed metadata for effective content management.

## 3.3 Nature and characteristics

Decisions about certain metadata characteristics vastly affect the limits and flexibility of what the metadata can offer within a domain. Content and metadata management requirements are also influenced by the topics presented here.

### 3.3.1 Granularity

In order to generate a structured metadata framework, boundaries of content, or an information object needs to be assessed. The term *information object* has been used to some extent in earlier literature. Gilliland-Swetland (1998) describes it as follows: *"anything that can be addressed and manipulated by a human or a system as a discrete*

*entity"*. An information object can in other words cover very specific content, but nothing prevents it from spanning over multiple items. No matter the scale or form of information objects, three topics can always be extracted: *content*, *context*, and *structure*. Content reflects what the information object deals with. Context is related to the creation of the object. Parameters such as place, time, reason etc. are connected to the context of an object. Structure defines the associations that are relevant to an information object, both internal and external.

The discussion above brings out the term *granularity*. What are the needs with regard to detail level? Should descriptive data concern individual, smaller parts of content separately, such as images and text of an article? Or should an article be treated as one package? *"Different kinds of metadata refer to the content essence at varying levels of granularity"* (Jokela 2001). In other words, metadata can be attached to the content at various detail levels depending on the boundaries set by the information object. Each information object can then additionally be described by multiple types of metadata, further enhancing the granularity. The greater the granularity, the greater the flexibility of the system, but at the same time management requirements are increased.

Mauthe & Thomas (2004, p. 80) state that metadata can be grouped according to the entity it describes. Object-related metadata refers to all metadata that describe an object as an entity. Segment-related metadata on the other hand describe parts of an object, or segments as the term implies, delimited by spatial or temporal boundaries. Examples of such segmentation include image regions and frames of video streams. This kind of segmentation enhances the granularity even more, as it takes a very specific information object and splits it into even smaller elements. De Stutter et al. (2006, p. 223) similarly divide this issue in half with a flat versus hierarchical metadata model approach. A broadcasted program might be considered as an elementary, indivisible unit. In a flat metadata model the descriptive data would be attached to the entire program, and individual parts of this unit could not be referred to. If the same program was to be modeled hierarchically, it would allow for detailed annotation of specific segments of the content.

Metadata with high granularity allows for very content specific queries. This in turn results in higher accuracy in search results. One clear disadvantage of high granularity is that it clearly requires a great deal of management efforts. A cost assessment based on long term needs is necessary when making decisions about these matters.

### 3.3.2 Free-form or controlled values

Metadata values can be created as free-form or they may be restricted by vocabularies, taxonomies, or ontologies. Alternatively, it may be a combination of these. *"Vocabularies are recommended lists of appropriate values, that define the value space of a data element"* (Duval 2001, p. 597). Duval mentions that higher semantic interoperability is achieved by limiting input to recommended values. As a result, such metadata will more likely be understood by others. On the other hand, if the vocabulary is not rich enough, it makes it hard to create sufficiently precise and accurate metadata. As a result, the content

and its description may not connect to each other with desired accuracy. This semantic gap can be described as *"the lack of coincidence between the information that one can extract from the… data and the interpretation that the same data have for a user in a given situation"* (Smeulders et al. 2000, p. 5). Smeulders et al. are referring to visual data, but the same applies to any form data if the interpretation and the actual content essence do not correlate.

A vocabulary should be constructed by forming some level of semantic metadata based on the usage area (Jokela 2001). The decisions concerning the vocabulary are important as the same term can have several meanings depending on the domain in which it is used. Hierarchical structures and ontologies can be very valuable in these cases. A combination of vocabulary and free form offers flexibility to some extent, while keeping the metadata structure under control. As an example, a newspaper may offer reviews of new movies. The rating element could be limited to numerical values between one and five. The genre element could have a fixed text pool consisting of 10 or 30 genres, depending on the agreed detail level. The title of the movie would however need to be in free-form.

There exist some examples of where free-form metadata has been implemented exclusively, without any vocabulary or grammar. These systems purely rely on tagging objects with one or multiple words in a single metadata field. Users have total control, and freedom, of what values they use for labeling the content. The web based communities Flickr and Deli.ciou.us are both prime examples of this kind of tagging (Mathes 2004). Both communities share a highly unrestricted approach to metadata, which could be the reason they have achieved such great popularity. It could be the simplicity of the systems that draw people's interest. Deli.ciou.us for instance shows the most popular tags directly on the front page (Mathes 2004). This could however encourage users to apply similar tags on their content, which in turn could increase the risk for overexploiting popular tags on irrelevant content. This would obviously reduce the relevance of search results.

Mathes (2004) also mentions that in addition to the tags shown on the front page, related tags are determined automatically by the system and shown on individual tag pages. One thing to note is that there is no hierarchical structure present in the system. Instead, there exist multiple implicit relationships between these terms, i.e. clusters of similar or related tags. This organic system that is developing has been called "folksonomy" by Thomas Vander Wal (according to Mathes 2004). Services like Flickr and Deli.ciou.us are interesting indeed, but for higher requirements on professional content management something more controllable is essential. However, even if more control is needed, tag clusters could be used in addition to some other stricter configuration.

## 3.4 Metadata types

There are numerous types of metadata to choose from, different ways of connecting it to content, and the purpose of using metadata also varies to a great extent. Metadata attribute types provide a broad definition of what metadata is used for, i.e. they describe the purpose of specific metadata within some domain. This has led to countless categorization efforts, which can be found in earlier literature. Words such as metadata *categorization*,

*classification*, and *types* are commonly used to divide metadata into different groups. Boll et al. (1998) identify three broad types of metadata, while Gilliland-Swetland (2000, p. 4) states more generally that *"it is helpful to break it [metadata] down into distinct categories… that reflect key aspects of metadata functionality"*. According to Jokela (2001) categorization can be based on usage of metadata, stages in the life-cycle of metadata, and characteristics of metadata. Boll et al. suggest a role-based categorization for classifying different kinds of metadata (1998; according to Jokela 2001). Jokela (2001) derived his own categorization based on their suggestion.

As can be seen, various viewpoints on how classification should be carried out exist. Here a summary covering these efforts is included. By comparing the purpose of categories formed by several authors (Boll et al. 1998, Gilliland-Swetland 2000, Jokela 2001, and Mauthe & Thomas 2004), a more abstract view of different metadata types could be found. Instead of presenting categorization efforts provided by different authors sequentially, similar levels of metadata abstractions are presented in parallel. Table 3 represents metadata categories found in literature especially suited for cross media publishing. These metadata types are used later on in this study, and are discussed in more detail in methodology Section 4.3.4 and results Section 5.2.

**Table 3. Metadata types based on categorization efforts from related literature**

| Metadata types | Type | Description | Example |
|---|---|---|---|
| **Essence** | **Content-specific** (Boll et al. 1998) | Metadata that is derived solely by content and independent of media type. | Text, images and video may all share the same content specific metadata. The content is simply presented with different media types. |
| | **Content-related** (Mauthe & Thomas 2004) | Describes the actual content or subject matter. | Description of video and image content. |
| | **Descriptive - Content-based semantic** (Jokela 2001) | Describes semantic qualities of the content essence answering the question what the content essence means. It is needed for the processing or usage of the content essence. It describes such qualities such as the subject, location, names, and style of the content essence. | Keywords of a news story. Also typically used in highly content-based products, such as in personalized news services, |
| | **Descriptive** (Gilliland-Swetland 2000) | Metadata used to describe or identify information resources. | Cataloging records, finding aids, specialized indexes, and hyperlinked relationships between resources. |
| **Contextual** | **Descriptive – Contextual** (Jokela 2001) | Describes the environment and conditions of content essence and its creation. | Geospatial information, timing information, and information on the equipment used to produce the content essence. |
| **Structure** | **Structural** (Jokela 2001) | Describes the structural characteristics, the format, of the content essence, but does not contain information about what the content essence actually means. | Decoding information related to e.g. video, audio, or graphics formats. Other relevant areas include compression data, composition and synchronization information, as well as information on sequencing the content essence. |
| | **Material-related** (Mauthe & Thomas 2004) | Describes available formats, encoding parameters, and recoding specific information. | Video or image format, their compression type and amount etc. |
| | **Media type-specific** (Boll et al. 1998) | Different media types have specific needs concerning metadata. | Texture of images, frequencies in audio, etc. |
| **Administrative** | **Control** (Jokela 2001) | Related to controlling the flow of content. It is used to determine whether the content is ready to progress in the content value chain or not. Commonly more temporary in nature, when compared to other metadata elements. | Machine control, QoS, error management, etc. |
| | **Media processing-specific** (Boll et al. 1998) | Describes functions for processing specific media. | Search and retrieval functions, functions for directing transfers, performance measuring to insure sufficient QoS, etc. |
| | **Administrative** (Gilliland-Swetland 2000) | Metadata used in managing and administering information resources. | Acquisition information, rights and reproduction tracking, documentation of legal access requirements, location information, version control and differentiation between similar information objects, and audit trails. |
| | **Use** (Gilliland-Swetland 2000) | Metadata related to the level and type of use of information resources. | Exhibit records, use and user tracking, content re-use and multi-versioning information. |
| | **Location-related** (Mauthe & Thomas 2004) | Describes location and amount of copies, condition of carriers, etc. | Could specify if there is a "master" version as well as copies made of it. |

One thing to note is that metadata categories do not by themselves specify anything about particular properties of implemented metadata fields. The original categorization efforts proposed by the authors mentioned in Table 3 are included in Appendices 12-15.

## 3.5 Standards for multimedia metadata

This section briefly presents metadata standards that are suitable for cross media and editorial processes. Using standards increases interoperability, but at the same time it sets boundaries on the system. Selecting between localized grammar and a standard should be done on a needs, cost, and management assessment basis. The level of semantics that metadata is capable of offering is partially dependent on the chosen grammar or standard. Compulsory or voluntary metadata elements are specified, into which the actual metadata is inserted. Localized grammar is many times preferred to large standards, mostly because it allows for a tailored system for domain specific needs. However, systems supporting standards are easier to configure for import/export of data. The next section covers metadata categorization, which is an effort for grouping similar metadata elements regardless of grammar or standard.

### 3.5.1 IPTC: NewsML-G2 and NewsCodes

IPTC stands for International Press Telecommunication Council. They provide a range of news exchange formats, as well as metadata taxonomies for the news industry. NewsML-G2 is one of the media-independent standards, which is especially well suited for exchanging general news content. It provides a means for describing the content, the management data, the packaging data, and also the transfer data. Bundling of multiple news items, such as articles, text, videos etc. is also supported. NewsML-G2 provides the grammar for the metadata, while NewsCodes provides a rich vocabulary consisting of standard terms for describing news. NewsML-G2 can be combined with the NewsCodes taxonomy. (Anon 2009b)

NewsCodes is split into multiple taxonomies, which are grouped into four main areas: Descriptive-, Administrative-, Transmission-, and Exchange Format NewsCodes. Taxonomies offer increased manageability as topics often relate to some specific area. Audiocodecs and Colorspace are two examples of vocabularies belonging to the Administrative NewsCodes. Each term in the vocabulary has four properties attached: (Anon a)

- The Concept Id: the full identifier for the concept of this entry expressed by a QCode
- The date when the entry was created
- A concept type of the entry, expressed as QCode and by a name in one or more languages
- The name and the definition in one or more languages

### 3.5.2 Adobe XMP

Adobe XMP is an extensible metadata platform that is built on the standards of the Semantic Web. The formal structure is adopted form the RDF standard, but XMP is implemented in XML for ease of parsing. Adobe XMP offers different ways of enhancing workflow productivity, but this requires the use of Adobe applications in the workflow. As an example, batch processing of metadata on image sets is possible. (Anon (b))

Adobe also presents revenue opportunities by using XMP. The community of advertisers could automatically be notified of a coming story, and a specialized application could auction ad space to accompany the web version of the story. (Anon (b))

**Table 4. Properties of Adobe XMP (Anon (b))**

| DOMAIN | WHO DETERMINES CATEGORIES? | WHO DETERMINES STRUCTURE? | WHO SETS VALUES OF CATEGORIES? | EXTENSIBLE? |
|---|---|---|---|---|
| Files created with Adobe products:<br>Adobe Creative Suite<br>Adobe InDesign®<br>Adobe InCopy®<br>Adobe Photoshop<br>Adobe Illustrator®<br>Adobe GoLive®<br>Adobe FrameMaker®<br>Adobe Acrobat® | User selects from several standard category sets supported in base XMP and/or implements own category sets | Adobe, based on the Resource Definition Framework (RDF) standard of the World Wide Web Consortium (W3C) | User | Explicit procedure provided by Adobe together with tools and training materials |

### 3.5.3 Dublin Core

The Dublin Core is a simple and effective way of describing a variety of networked resources. The goals of the Dublin Core are simplicity of creation and maintenance, commonly understood semantics, international scope, and extensibility. The standard consists of two main levels, and it is up to the user to choose which one of them that is more suitable for a given task. The Simple Dublin Core consists of fifteen basic elements, while the Qualified Dublin Core includes three additional elements and a group of element refinements for improved element semantics. The fifteen basic elements are presented in Table 5. Each element of the Dublin Core is optional and may be used multiple times. The order of elements is free, but the provider may e.g. use this freedom to provide significance for the sequencing of multiple occurrences of the same element. (Hillmann 2005)

**Table 5. Simple Dublin Core element set (Anon 2008b)**

| Element | Description |
|---|---|
| **Title** | The name given to the resource. |
| **Subject** | The topic of the content of the resource. |
| **Description** | An account of the content of the resource. |
| **Type** | The nature or genre of the content of the resource. |
| **Source** | A reference to a resource from which the present resource is derived. |
| **Relation** | A reference to a related resource. |
| **Coverage** | The extent or scope of the content of the resource. |
| **Creator** | An entity primarily responsible for making the content of the resource. |
| **Publisher** | The entity responsible for making the resource available. |
| **Contributor** | An entity responsible for making contributions to the content of the resource. |
| **Rights** | Information about rights held in and over the resource. |
| **Date** | A date associated with an event in the life cycle of the resource. |
| **Format** | The physical or digital manifestation of the resource. |
| **Identifier** | An unambiguous reference to the resource within a given context. |
| **Language** | A language of the intellectual content of the resource. |

### 3.5.4 IPTC: Photo Metadata

IPTC Photo Metadata provides data about photographs, and each metadata entity is defined as a property. The metadata entities are grouped into administrative, descriptive and rights related properties. IPTC Photo Metadata does not cover technical metadata produced by cameras, such as ICC profiles and positioning data. The core schema consists of 32 metadata properties, but the standard offers extensions. The core schema can be found in Appendix 16. (Anon 2008c)

### 3.5.5 DIG35

The DIG35 standard aims to provide a standardized mechanism which allows end-users to see digital imaging as equally convenient and flexible as traditional photographic methods, while offering benefits that were not possible using traditional methods. The key focus areas of the DIG35 Initiative Group include (Anon (c)):

- Defining a standard set of metadata for digital images that can be widely implemented across multiple image file formats
- Providing a uniform underlying construct to support interoperability of metadata between various digital imaging devices
- Ensuring that the metadata structure provides both a common inter-application exchange format and a high-degree of extensibility for enhanced use by specific applications

### 3.5.6 EXIF

The EXIF standard is applicable to image and sound content, and is especially designed for use in digital still cameras. The audio format specification is not discussed here. The image file specification does not only specify structures for metadata. It contains the following specifications:

- Format version definition

- Specifications relating to image data
- Basic structure of image data
- Tags
- JPEG marker segments
- Specifications on data recording

The EXIF standard offers a variety of different tags. It includes tags related to:

- Image data structure (width, height, orientation…)
- Image description, artist, copyright…
- Version
- Image configuration
- User comments
- Date and Time
- Picture-taking conditions
- GPS

(Anon 2002)

## 3.6 Relationship between metadata and content management

The amount of digital content that is created by current media companies and the pace at which it is produced is astounding. Without proper ways of handling this content its full potential remains unexploited. Gilliland-Swetland (2000, p. 9-11) proposes improved metadata for maximizing use of assets in multiple areas:

- *Increased accessibility* by the existence of rich and consistent metadata.
- *Retention of context* for documenting and maintaining complex relationships and associations among collections of objects, people, places, movements, and events. Also for indicating the authenticity, structural and procedural integrity, and degree of completeness of content.
- *Expanding use* – in order to attract and reach new users current content can be modified to suit different needs.
- *Multi-versioning* for tracking similarities and differences between different versions of content.
- *Legal issues* can efficiently be attached to content via metadata, which aids the tracking of rights and reproduction information.
- *Preservation* – metadata is essential for enabling the existence of system independent content, which will survive technological updates.
- *System improvement and economics* - benchmarking of technical data allows for the construction of more effective and efficient systems.

The definition of the term *content* was discussed in Section 2.2. According to that definition content cannot exist without essence and metadata. This is taken even further by the following definition of the term *asset* (SMPTE; according to De Stutter et al. 2006 p. 221):

*asset = content + right to use it*

According to this, company assets only exist when the essence, metadata, and legal rights are known. The increased need for careful documentation of rights and protection information is also mentioned by Pereira et al (2008, p.731). Furthermore, metadata is considered to play an integral part in connecting defragmated organizational wide resources (Anon 2009a). Topics discussed in Section 2.2 are also relevant to the theme of this discussion.

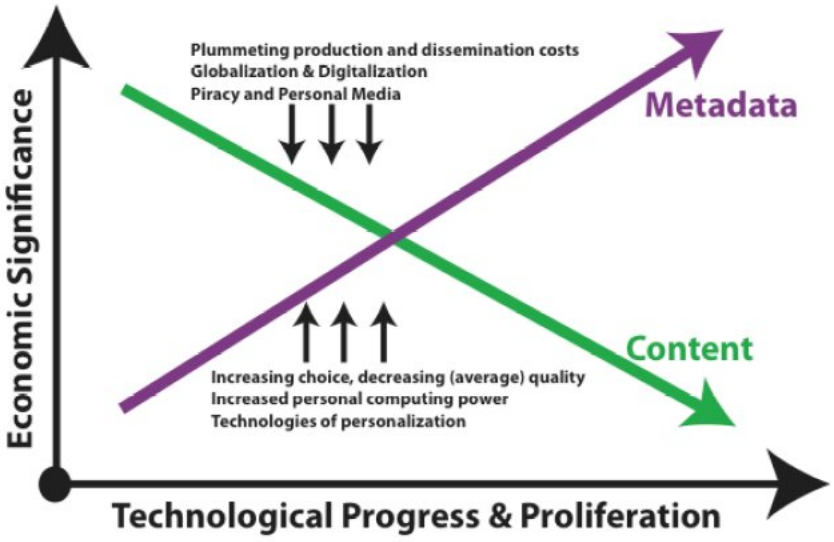



**Figure 9. The metadata twist (Haase 2004, p. 205)**

Figure 9 above represents *the metadata twist* presented by Haase (2004). He states that economic significance of metadata increases as the amount of available content increases. At the same time the average value of content decreases. This will lead to a point in time when the value of descriptive metadata will become more valuable than the content it describes. Haase (2004) also emphasizes the importance of high-quality precise metadata. In conclusion, well placed metadata is a good investment:

> *"Metadata is like interest – it accrues over time."*
> (Gilliland-Swetland 2000)

# 4 METHODOLOGY

## 4.1 Methodology for studying metadata in editorial systems

### 4.1.1 Selection of methods and data

Suitable empirical research methods should be chosen based on the research questions and goals. This is to some extent possible, but due to the nature of qualitative research several aspects can emerge during the study. Qualitative research is emergent rather than tightly prefigured (Creswell 2003, p. 181). Empirical qualitative research takes place in the natural setting, and is therefore a highly involved actual experience: *"the researcher filters the data through a personal lens … one cannot escape the personal interpretation brought to qualitative data analysis"* (Creswell 2003, p. 181-182). This is true to some extent, but with well designed methods subjective interpretation is kept to a minimum. Creswell also states that the idea behind a qualitative study is to select participants, sites, documents, or other material, which are optimal in the sense that they are carefully chosen to support the researcher in answering the research questions (2003, p. 185).

The study presented in this thesis deals with editorial processes in media companies. This involves a generous amount of actors, processes, different content types and delivery media. The analysis methods therefore need to assist in dividing the complex analysis and communication task into smaller manageable subtasks (Salminen et al. 1996, p. 76). According to Creswell, Miles and Huberman (2003, p. 185) identified four aspects of qualitative research which included the setting, actors, events, and processes, which is exactly what is included in this study.

Related literature was examined in order to see if a suitable framework could be found to support the data gathering process. Archived data samples could be used exclusively as an information source for metadata, but including qualitative research was essential in order to answer the research questions. The data gathering, research methodology, and modeling techniques that were used in this thesis are discussed in more detail in Section 4.3.

### 4.1.2 Methodology for studying legislative documents: RASKE project

Related literature did not seem to offer any methods that would have suited this study directly. However, while gathering information about document analysis, one project which had faced similar challenges was discovered. The RASKE project had been active between years 1994-1998, and had developed methods for studying legislative documents as well as actors, activities, and systems involved in the creation and handling of these. The developed methods consist of various analysis and modeling methods, which have been tailored and developed for the purpose of document analysis. From this methodology, appropriate parts have been selected and used in the research's empirical part.

Salminen et al. (1997, p. 645) give a short overview of the RASKE project. The project was launched in spring 1994 by the Finnish parliament and a software company, in corporation with researchers at the University of Jyväskylä. It was motivated by document

management problems in the Finnish parliament and government. The goal was to analyze legislative documents, develop new standards for these, and study and develop methods for the standardization. The term RASKE is derived from the Finnish words *"Rakenteisten AsiakirjaStandardien KEhittäminen"*, which stands for the development of standards for structured documents. Lyytikäinen (2004, p. 27) states that the major practical purpose of the project was to enhance document management and accessibility of the information created in the Finnish Parliament and ministries.

Although this thesis focuses on metadata, the challenges and goals of this thesis and the ones of the RASKE project share some important similarities. Salminen et al. (1996, p. 72) mention several focus areas for electronic document management strategies in their paper: *"Standardization may be regarded as a cornerstone of [electronic document management] strategy. …For defining effective standards for digital documents of an organization, a profound document analysis is needed. In the analysis, current documents and document management practices are studied and described, and new document structures and document management practices are developed."* The environment of electronic document management in enterprises, consisting of activities, actors, systems, and documents, has been defined by the *"RASKE model for electronic document management environments"*, which is illustrated in Figure 10 (Lyytikäinen 2004, p. 30).

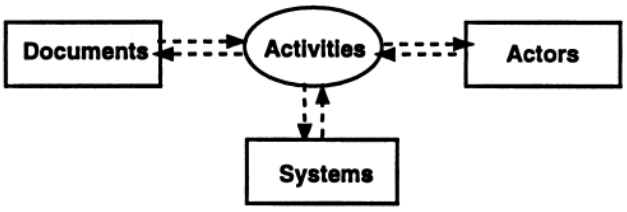



**Figure 10. The RASKE model for electronic document management environments (Lyytikäinen 2004, p. 30)**

*Documents* refer to recorded data that is intended for human perception. Documents can be identified and handled as a unit in the activities. *Systems* are an abstraction for hardware, software, or anything else that is needed in the management of the documents. The editorial software that were used in the case companies are not presented in more detail in this study. An overview of editorial software used in Sanoma Magazines can be found in Alanko (2009). It is important to separate documents from systems, as documents should be available after system changes. *Actors* are people and organizations performing activities and using documents as well as systems in the activities. (Salminen et al. 2000a, p. 625)

Salminen et al. (1997, p. 645) note that some of the problems identified prior to the RASKE project by teams studying the legislative work in the parliament concerning document management included:

- Incompatibilities of the systems cause need for repeated typing of the same text. This in turn causes inconsistencies in documents as well as a demand for additional labor.
- The retrieval techniques of different systems are heterogeneous.
- The retrieval techniques of the electronic archiving system and the tracking system are not satisfactory.

All of above mentioned challenges have also been identified in the CrossMedia project, prior to the actual research. The means of cross media publishing are clearly different from legislative work. However, one cannot deny the similarities when strictly seen from a document management perspective. Salminen et al. (1997, p. 645) write: *"In spite of the methodology that is developed for a special case – to solve problems in the management of legislative documents in Finland – the methods themselves are not application dependent"*.

As the reader may have noticed, the RASKE project continuously refers to document structures and document analysis, not mentioning metadata. Further investigation about the RASKE project revealed that RASKE2 was launched a few years later with the goal of developing methods for the integration of information resources by means of metadata standardization. RASKE2 builds upon discoveries made in RASKE, and the methodology formed in RASKE served as a basis for RASKE2. Further discussion about the RASKE and RASKE2 projects can be found online at (Anon 2009a).

## 4.2 Research sites

Three media companies participating in the CrossMedia project were selected as research sites for this study. These consisted of a newspaper, a magazine publisher, and a news agency. Due to the limited amount of time available and the needed amount of visits per company, five separate research sites within these companies were decided to be sufficient. These five research sites consisted of the editorial office at a weekly supplement at the newspaper, three separate magazines and their respective editorial offices from the magazine publisher, and the editorial office at the news agency. The research sites were selected in such a manner, that they would form a group containing sufficiently diverse editorial processes. Contact persons and specialists in their respective area were selected from each participating company. These key persons then assisted getting access to the sites. Each site involved in this study is presented in the following sections.

It should be noted that one of the magazines, MB, was left out of the results excluding the modeling of processes, as a traditional editorial system was not used in their editorial processes. Furthermore, results concerning image metadata are not taken into account for the news agency, as observed workflows did not include image content.

### 4.2.1 Helsingin Sanomat

Sanoma is the leading newspaper publisher in Finland. In addition to Helsingin Sanomat (hereafter shortened as HS), the largest subscription-based daily in the Nordic region, Sanoma publishes national and regional newspapers, local papers, free sheets, and provides digital services. Sanoma also provides corporate customers with business information,

photo agency, and news analysis and summary services. Sanoma is part of the SanomaWSOY Group. (Anon 2008a)

Main focus during the modeling of editorial processes was on the workflow in NYT supplement. Questionnaires were also handed out at this site.

### 4.2.2 Sanoma Magazines Finland

Sanoma Magazines Finland (hereafter shortened as SM) is Finland's leading magazine publisher and the market leader in women's titles, family magazines, children's magazines, youth titles, and consumer IT publications. They are also strong in custom publishing and are devoting increasing resources to online services. Altogether Sanoma Magazine publishes over 40 consumer titles. Sanoma Magazines Finland is part of Sanoma Magazines division and SanomaWSOY Group. (Anon 2008a)

Three different magazines were visited from Sanoma Magazine's Finnish division. These included ET, MikroBitti (shortened as MB), and Custom Publishing (hereafter shortened as CP). ET and MB can both be regarded as regular magazines, but with dissimilar target audiences. ET is a women's title / family magazine, while MB focuses on consumer IT. Custom publishing on the other hand differs to a great extent from the other magazines in the way content is produced.

### 4.2.3 STT

The Finnish News Agency (STT) is an independent, national news provider. STT is the only news organization in Finland that produces a real-time and comprehensive news agency service. STT also produces a number of other services for media as well as communications services for leading companies and other organizations. (Anon 2008a)

## 4.3 Analysis – Methods and data

### 4.3.1 Overview

The case studies of this research consisted of two rounds. The first round focused on specifying the domain, process modeling, document modeling, and role modeling. User needs, i.e. the needs of the people involved in the content creation process at the editorial office, were revealed to some extent during the observations. This step supported the data gathering required to answer research questions 1, 2, and 3. During the second round, models created based on the first round were enhanced. The main focus of the second round was on direct input from actors in the form of interviews and questionnaires. User needs were analyzed, but also other viewpoints were discussed to further enhance the conception of the domain in question. This phase therefore aided in answering research questions 1, 2, and 4. An overview of the entire analysis process is presented in Figure 11.

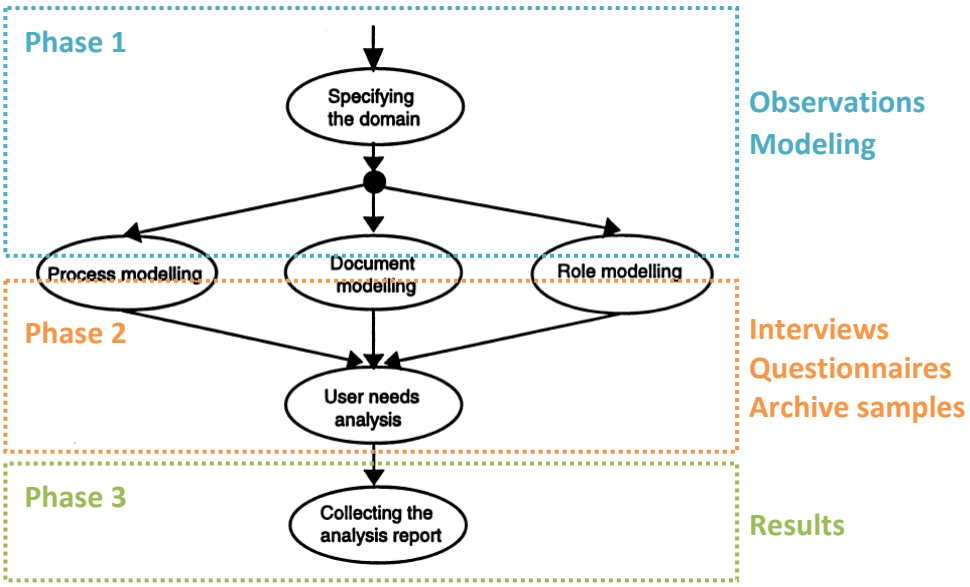



**Figure 11. A model for the document analysis process (Salminen et al. 2000a, p. 630)**

The entire analysis consisted of three distinct phases. Phase one began by *defining and describing the domain* of interest. In other words, the boundaries for the study were set, and the restricted domain could then be examined more closely. Large and complex domains may be divided into subsystems, where after they are treated exactly in the same manner (Shlaer & Mellor 1992; according to Salminen et al. 1996, p. 76). In this study this was done to some extent, because focus was set on the editorial processes of the participating media companies. This is a small subset of the entire domain of each company. However, the focus area of this study could also be seen as the core domain, because no further interest is given to any other parts of the media companies. Once the domain definition was completed, the analysis proceeded with *process*, *document*, and *role modeling*, as suggested by Salminen (2000b, p. 305-306). These can be executed one at a time, but due to their nature they cause iteration and are often executed in parallel, which was the case in this study. In *process modeling*, smaller activities of the domain were identified, as well as the organizations responsible for them and the documents created or used in these activities. The *document modeling* on the other hand described document types, their life cycles, their content, and their relationships to each other. *Role modeling* identified the most essential document users, where after their document management activities were described. These tasks were performed during the first round of visits at the case companies.

The second phase of the analysis continued by enhancing the models made during the first round of visits. This phase therefore included new visits to the research sites. In addition to enhancing the data from the first round, the second round also included two types of surveys: interviews and questionnaires. The participating media companies were also asked to deliver metadata archive samples at this point. The specifications for the archive samples can be found in Appendix 2. The *user analysis* focused on the needs concerning future document management (Salminen 2000b, p. 305-306). Even though the model in Figure 11 suggests that user analysis is separated from the three main analysis processes, it

was natural to also take these aspects into account during the process, document, and role modeling.

The third phase of the analysis process consisted of collecting the analysis report. During this stage final examination and analysis of collected data was performed. The outcome is presented in the results section of this study.

The document analysis model presented in Figure 11 is a good abstraction of the analysis processes performed in this study, but the actual analyses within each step differ from those of the RASKE project. For a more comprehensive discussion of the other RASKE modeling methods, the reader is suggested to have a look at the work of Salminen (2000b), Salminen et al. (1996, 1997, 2000a), and Lyytikäinen (2004). The methods used for modeling the editorial processes and metadata flow are presented in Section 0. Construction of the metadata framework is discussed in Section 4.3.4.

### 4.3.2 Observations

According to Creswell, multiple methods are available for qualitative research, many of which highly interactive and humanistic (2003, p. 181-182). This includes observation, interviews, and questionnaires among others methods. These three were utilized in this study, as they were assessed to be good candidates for harvesting the data needed to answer the research questions.

Observations allow extraction of data, which is more likely to represent a more objective view of the workflow, documents used, and other related matters than the data an interview alone would produce. The observations were vital for constructing and forming questions for the interviews and questionnaire, and were thus done prior to these. Without the knowledge gained through observations the questions could easily have been focused on irrelevant matters. In order to collect as much data as possible, audio was recorded during the observations. This ensured that any unclear or missed details could be extracted at a later point. During the observations 14.5 hours of audio material was recorded. The persons who participated in the observations were selected in such a way that they formed a broad selection of actors within current editorial processes (see Table 6). Some of the involved people were chosen by the companies.

**Table 6. Actors from the case companies that participated in the observations**

| Company | Department | n | Work description |
|---|---|---|---|
| **STT** | International department | 1 | Journalist / sub-editor |
| | Online department | 1 | Journalist / sub-editor |
| | Homeland and reporter department | 1 | Journalist |
| **HS** | Archive department | 1 | Archive manager |
| | Photo department | 1 | Photo journalist |
| | NYT - supplement | 2 | Journalist |
| | City department | 1 | Journalist |
| **SM** | ET | 1 | AD |
| | | 1 | Sub-editor |
| | MB | 1 | Managing editor |
| | Custom Publishing | 1 | Producer |
| | | 1 | Graphic artist |

The documents collected during the observations consisted of screenshots of editorial software and metadata fields, files containing metadata from the system, software instructions, instructions for inserting metadata, and various other types of documents concerning the editorial system or metadata. These documents were then used to complement the information gathered during the observations, as all data could not effectively be collected in real time. Most of the documents were collected during the first round of visits to the companies, but also to some extent during the second round. Altogether 1269 documents were collected during the field case studies.

Different properties of metadata fields were documented in addition to field names and values. Some of the metadata field properties discussed below could directly be determined by examining the metadata fields, while assessment of a few field properties required feedback from the user. As mentioned earlier, qualitative research need not be tightly prefigured, which was especially true in this case as new field properties were discovered during the observations. These discoveries introduced a need for iterative procedures, as the already examined fields needed to be updated with the newly discovered properties. The final list of metadata field properties that were documented in the metadata inventory, based on the iterative process discussed, include:

1. Value insertion format
2. Insertion type
3. Dynamic nature
4. Multi-selection
5. Locked
6. Relationships
7. Special properties
8. Description
9. Additional information
10. Use

Value insertion format specifies in which form the value is inserted into the metadata field. The metadata field may only accept fixed values or it might accept text in free-form. Insertion type on the other hand defines the amount of automation that is applied for value insertion. The nature of the fields is related to the dynamics and value modifications that a field value goes through during the editorial processes. Multi-selection specifies if single fields accept multiple value entries. Locked fields allow no modification on the value. Relationships describe if two fields are connected or linked to each other. Special field properties were documented when possible, such as color coding. Description of fields as well as additional information was also collected when possible. During the observations users were asked to indicate which metadata fields they use, and also estimate the usage of those fields. Estimation with high accuracy would have been difficult, which is why users were asked to indicate usage by selecting one of the groups shown in Table 7 for each metadata field that they were familiar with in some way. Some semi-automated or fully-automated fields were marked as always used, even though the user did not interact with

the fields in anyway. Fields which the users did not know the meaning of were marked as "n/a".

**Table 7. Groups indicating metadata usage frequency**

| | |
|---|---|
| **Never** | This field is not in active use (may have been used in the past). |
| **Occasionally** | This field is used occasionally. Usage might be highly individual or e.g. related to content. |
| **Always** | This field is almost constantly used. Only left blank because of special conditions or errors. |
| **n/a** | The user did not know what the purpose of the field was |

### 4.3.3 Modeling

BPMN (Business Process Modeling Notation) (Anon 2009c) is a standard for modeling business processes, which aims to provide a notation that is readily understandable by a wide range of business users, from technical developers to people who monitor the processes in action. In other words, BPMN strives to create a standardized bridge between the gap of business process design and business process implementation. BPMN is flexible by supporting extensions to BPMN diagrams, such as new markers and indicators. New shapes, or elements, may also be added as long as they do not conflict with an existing shape specified for BPMN. The use of coloring is also free, and as such offers a way to further extend information conveyed by elements.

Based on these properties, BPMN was chosen as the modeling notation for modeling the editorial processes of the case companies. Other more complex modeling languages such as UML could have been used, but BPMN offered all features that were needed in this case. It provided means for modeling user workflows separately, but at the same time allowed for interaction notation between them. Flows between activities and documents could also be annotated. BPMN supports separate flow types, which was found useful when modeling the editorial processes. As an example, it allowed for clear visualization of when information manually needed to be exchanged between actors instead of automatically through the editorial system. BPMN allows for annotation of sequence flows, message flows, and associations.

The goal of the modeling was to identify similarities and differences between the editorial processes of the participating media companies, as well as to show how metadata evolves alongside the content creation in the systems. The models identify the actors, processes, documents, and metadata on an abstract level. Based on the models discussed in Section 2.3, the editorial process was divided into four main steps: planning, production, publishing and distribution, and archiving. This gives an overview of how the content is created and handled at various key-stages in the content creation process. Publishing channels are also illustrated. To better visualize these sub-processes, each step was colour coded in the models as follows:

- Planning – yellow
- Production – blue
- Publishing and distribution – green
- Archiving – grey

Due to differentiated metadata practices in the research sites the metadata time-cycle was modeled separately from the editorial processes. This gave a better and clearer understanding of the current flow of metadata and content within the editorial processes.

### 4.3.4 Metadata framework

The creation of a metadata inventory, based on the current metadata structure of the case companies, was an important step in this study. This metadata inventory was then further abstracted into a metadata framework, which defines essential metadata field attributes of current editorial systems. Based on the findings of this study, suggestions for future attributes and improvements on current attributes are presented in the conclusions section.

The data used for creating the metadata framework consisted of metadata fields collected during observations at the research sites. The companies were also requested to supply samples (n=200 / company) from their archives containing metadata fields and values for both images and text material. These samples made it possible to identify the importance and use of specific metadata fields. The specifications for the archive samples can be found in Appendix 2. The archive samples were not completely randomized according to the given specifications. The samples were often chosen from a few specific days, when they preferably should have been selected randomly from a time-period of one year (2008) as noted in the specifications. A time range of one year was decided to be long enough to contain diversified data. A longer time span would have been possible, but risks of system changes within the time span increase. STT's archive samples concerning images were not used in the analysis, because none of the observed processes at STT concerned images or photos. Altogether 800 documents were examined regarding metadata. Qualitative and quantitative data were compared to each other.

The metadata fields that were discovered during observations were first analyzed separately on a media company level. Fields that were discovered during the observations, but which proved unimportant based on the data found in archive samples were discarded. The remaining metadata fields with similar functional or semantic purposes were then clustered together by following the prioritization in Table 8. These clusters could then be compared side by side, unaffected by software and other company specific boundaries. This in turn generated new sub-clusters that were used to form a logical structure for the metadata framework. The goal of the framework is not to serve as a direct mapping scheme between the different metadata structures. It is an abstraction of similar metadata that was used within the editorial processes of the different media companies. Some of the metadata field attributes suggested in the framework are not based on combinations of metadata from all current metadata structures. These were included in the framework as they were seen as good candidates for future editorial purposes.

The main clusters were matched to a high-level division of four types of metadata (see Table 3), distilled from previous literature (Boll et al. 1998; Gilliland-Swetland 2000; Jokela 2001; Mauthe & Thomas 2004). This allowed for logical processing of the combined metadata collected from the research sites, and ultimately led to the creation of the metadata framework. The used metadata types consist of essence, structural, administrative, and contextual metadata. The validity of these metadata types were confirmed by the data that was collected at the research sites. Altogether 430 distinct fields were assigned to these categories, of which 301 fields were used for the metadata framework.

The generation of the metadata framework thus consisted of the following phases:

1. Creating records of discovered metadata field names in editorial systems during observations.
2. Assessment of what the fields were used for.
3. Grouping of the metadata fields according to the metadata types derived from literature.
4. Identification of similarities between metadata fields within the newly formed groups. Forming of sub-clusters based on findings.
5. Generation of common metadata field attributes based on sub-clusters.

The assessment of metadata fields described in step 2 prioritized values entered by users. This was the most accurate way of extracting the true use of specific fields. If values were not available, then prioritization was done according to the order in Table 8. User descriptions of fields were collected during observations. The participating companies sent field descriptions to accompany the delivered archive samples, but the descriptions did not cover all the discovered metadata fields. If none of the above mentioned information was available, then assessment was based on the name of the metadata field.

**Table 8. Prioritization used for determining the purpose of metadata fields**

| | |
|---|---|
| 1. | Field values entered by users |
| 2. | User description |
| 3. | Field description provided by participating companies |
| 4. | Field name in system |

Construction of the framework began when all the fields had been collected and assessed based on their purpose. Table 9 shows a small part of the metadata framework in an early stage. The fields are only organized with respect to the editorial software used and media company.

**Table 9. Initial metadata inventory, only organized according to software and company Phase 1 & 2**

| Organization | System | Field name | Finnish field name |
|---|---|---|---|
| HS | MediaKsi | keywords | asiasanat |
| | | free keywords | vapaat asias. |
| | | continents | maanosat |
| | | countries | maat |
| | | city | kaupunki |
| | | city areas | kaupunginosat |
| | | characters | merkit |
| | CCI | length | pituus |
| STT | Doris Job | address | osoite |
| | | district | paikkakunta |
| | | content categorization | aihesanaluokitus |
| | Doris | area | alue område |
| | | content categorization | asialuokitus |
| | | SMS-selections | SMS-valinnat |
| | | online department | onlineosasto |
| | | telegram | sähke |
| | | XML | XML |
| | | SMS | SMS |
| | | main news | pääuutiset |
| SM - YJ | Doris | name | nimi |
| | | id-number | id-numero |
| | | name | nimi |
| | | page pre-set | sivupohja |
| SM - ET | Doris | additional information | lisätiedot |
| | | additional information | lisätiedot |
| | | story pre-set | juttumalli |
| | | page pre-set | sivupohja |

The next phases (3 and 4) consisted of collecting metadata fields with similar purpose. The table below represents similar administrative metadata that is grouped together from different editorial systems. These fields were then examined and fields that contained very company specific data were dropped in order to obtain a selection that was suitable for all companies. By looking at the fields clear similarities could be found, which are highlighted in Table 10.

**Table 10. Similar administrative metadata fields highlighted by the same colour Phase 3 & 4**

| Organization | System | Field name | Finnish field name |
|---|---|---|---|
| STT | Doris | send out:date | uloslähetysaika:date |
| | | send out:time | uloslähetysaika:time |
| | | status | status |
| | | desk | deski |
| | | department | osasto |
| | | embargo:date | embargo:date |
| | | embargo:time | embargo:time |
| | Doris Job | status | status |
| | | story ready | juttu valmis |
| HS | CCI | copyright | copyright |
| | | production status | tuotantotila |
| | | work phase | työvaihe |
| | | production phase | tuotantovaihe |
| | | layout status | taittotila |
| | | publish date | julk.pvm |
| | MediaKsi | date | pvm |
| | | copyright | oikeudet |
| SM - YJ | Doris | status | status |
| | | archiving | arkistointi |
| | | status | status |
| | | publish date | julk. pvm |
| | | publish date | julk.pvm |
| SM - ET | | status | status |
| | | archiving | arkistointi |
| | | status | status |
| | | publish date | julk. pvm |
| | | publish date | julk. pvm |

By clustering similar fields together, more abstract fields could be identified and generated. This phase of the framework creation process is shown in Table 11, which includes status, control, and copyright metadata field attributes. The different sub-clusters create the basis for the metadata framework, which is presented in results Section 5.2. The field attribute types of the framework are also used for presenting the metadata time-cycle in results Section 5.1.5.

**Table 11. Generation of common metadata field attributes based on sub-clusters Phase 5**

| Metadata field attribute | Organization | System | Field name | Finnish field name |
|---|---|---|---|---|
| **Status** | STT | Doris | status | status |
| | | | desk | deski |
| | | | department | osasto |
| | | Doris Job | status | status |
| | | | story ready | juttu valmis |
| | HS | CCI | production status | tuotantotila |
| | | | work phase | työvaihe |
| | | | production phase | tuotantovaihe |
| | | | layout status | taittotila |
| | SM - YJ | Doris | status | status |
| | | | archiving | arkistointi |
| | SM - ET | | status | status |
| | | | status | status |
| | | | archiving | arkistointi |
| | | | status | status |
| **Control (date & time)** | STT | Doris | send out:date | uloslähetysaika:date |
| | | | send out:time | uloslähetysaika:time |
| | | | embargo:date | embargo:date |
| | | | embargo:time | embargo:time |
| | HS | MediaKsi | date | pvm |
| | | CCI | publish date | julk.pvm |
| | SM - YJ | Doris | publish date | julk. pvm |
| | | | publish date | julk.pvm |
| | SM - ET | | publish date | julk. pvm |
| | | | publish date | julk. pvm |
| **Copyright** | HS | MediaKsi | copyright | oikeudet |
| | | CCI | copyright | copyright |

### 4.3.5 Interviews

The interviews were performed during the second round of the field case studies. The questions were designed to support the data needed for answering the research questions. The questions were also partly formed based on findings from the observations in the first round of the field case visits. The questions were not fixed, but followed a pre-designed layout so that the same issues were covered during all interviews. The answers were meant to complement the results from the observations and the questionnaires. Thematic interviews with 8 people (Table 12) were conducted. They lasted on average 30 min. each and were audio taped. The persons that were interviewed during the second field case study had all participated in the first round's observations, with an exception for the sub-editor of MB. The interviews were performed in Finnish.

**Table 12. Actors from the participating media companies that were interviewed**

| Company | Department | Work description |
|---|---|---|
| **STT** | International department | Journalist / sub-editor |
| | Online department | Journalist / sub-editor |
| **HS** | Archive department | Archive manager |
| | Photo department | Photo journalist |
| | NYT - supplement | Journalist |
| **Sanoma Magazine** | ET | Sub-editor |
| | MB | Sub-editor |
| | Custom Publishing | Producer |

A checklist was used during the interviews to ensure that each participant gave relevant answers to the themes of interest. The interview focused on the following themes:

- Participant's relationship to the research topic
- Creation of metadata
  - Use of already existing metadata
  - Media types to which metadata is attached
  - Knowledge of currently available metadata fields
  - Attitude towards adding/modifying metadata
  - Time spent on adding/modifying metadata
- Use of metadata during content production and –management
  - When and how
  - Knowledge of who uses the created metadata
  - Search functionalities
- Metadata needs, difficulties, and proposals for improvement
  - Difficulties related to insertion or creation of metadata
  - Need for communicating outside the editorial system
  - Suggestions for improving the creation of metadata

### 4.3.6 Questionnaire

A questionnaire was created and distributed at the case companies' editorial offices. Both interviews and questionnaires complemented the findings from the observations, and the questions were therefore to some extent based on the knowledge gained from prior visits to the research sites. While interviews focused on other matters that needed more freedom for answering, the questionnaire allowed for specific questions that could be answered quickly. For example, the questionnaires allowed for assessment of the current use of metadata, but suggestions for future improvements are difficult to collect by using pre-defined questions. The questionnaire was designed not to be too long in order to encourage people to participate in the survey. A Likert scale from 1 (strongly disagree) to 5 (strongly agree) was used as it was decided to give a sufficiently wide scale for the results, while still being simple enough to fill in. A "cannot tell" option was also provided for each question in order to avoid forced answers. The questionnaire data was treated anonymously with an exception for the answers from those who also participated in the interview. Permission to

use their answers for cross-examination between questionnaire and interview data was sought. All participants agreed on having their answers treated this way.

The questionnaire was written in Finnish, contained 32 questions (see Appendix 3), and these questions were divided into four main groups. The first group of questions covered usage of metadata according to document type. The second group focused on the creation and insertion of metadata values. After this participants were questioned about the benefits of metadata on both a personal and organizational level, as well as their viewpoint on how the end-users benefit from metadata. Finally the participants answered questions about problems and restrictions of current metadata in editorial systems. All participants were also asked to specify their work description as background information. In total, 81 replies were received from the case companies. Three of the replies were from staff working for IltaSanomat (IS), which is a daily paper whose editorial office is situated in the same building as HS. A detailed overview of rejected, missing, and "cannot tell" answers are available in Appendix 5.

Missing and "cannot tell" answers equal to or exceed 20% in questions 5, 14, 16, 19, 24 and 25. Most of these peaks were explained by the work descriptions of participants and their answers on questions 1-5 regarding metadata usage frequency. For example, question 14 inquired about the usefulness of image metadata. The combined missing answers and "cannot tell" percentage dropped to 3% when participants that did not use image metadata in their work were filtered out based on answers to question 3.

Table 13 shows the number of work descriptions given by the participants of the survey. One thing to note is that the total amount of work descriptions exceeds the amount of participants. This is due to the fact that some of the participants had multiple work descriptions, and their work tasks may therefore vary to some extent.

**Table 13. Roles of respondents**

| | | Text content | | | | Image content | | | | | Managers | | Archive and technical | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Paper / Organization | | Journalist | Sub-Editor | Producer | Editor | Graphic designer | AD | Layout designer | Multimedia designer | Photo journalist | Archive worker | Technical specialist | Managing editor | Manager | Director of development |
| STT | | 14 | 5 | 1 | | 2 | | | | | | | | 1 | |
| HS | | 18 | 2 | 4 | 1 | | | 1 | 1 | 2 | 6 | | | | |
| SM | CP | | | 2 | | 4 | | | | | | | | | |
| | ET | 3 | 2 | 2 | | 2 | 1 | 1 | | | | | | | |
| | MB | 2 | 1 | | | | 1 | | | | | 2 | 1 | | |
| IS | | 1 | | | | | | | | | 1 | | | | 1 |
| tot. | | 38 | 10 | 9 | 1 | 8 | 2 | 2 | 1 | 2 | 7 | 2 | 1 | 1 | 1 |

Groups were formed based on work descriptions on order to more effectively process the data. Participants were classified into four groups (text content, image content, managers, and archive and technical) based on similar work descriptions. For the largest group (text content) the similarity among subgroups (journalist, sub-editor, producer, and editor) was verified by t-tests. These were combined into one large group. A maximum of one question with significant difference between the groups was set as a criterion for combining the subgroups. Similar groups are highlighted in Table 13.

Other groups of interest were formed based on metadata usage frequency, organization, and separation of people working with image and text content. Based on the work descriptions it was also possible to resolve if participants dealt with web related matters in their work. The formed groups are presented in Table 14.

**Table 14. Groups formed for analysis of questionnaire results**

| Independent variables | Values (n) |
|---|---|
| Work groups | Text content(51), Image content(15), Managers(3), Archive and technical(9) |
| Use frequency | Low(20), Medium(26), High(34) |
| Organization | STT(23), HS(31), SM – CP(6), SM – ET(11), SM – MB(7), IS(3) |
| Text vs. Images | High use of both(10), High use of images(15), High use of text(33), None(21) |
| Web related work | Produces content for web(12), Does not produce content for web(68) |

Use frequency was determined by the mean of questions 1-4 in the questionnaire, which are related to metadata usage frequency. Question 5 was disregarded because of the high amount of missing answers even though it was related to usage frequency. Users were split into three use frequency groups based on the mean of questions 1-4 as follows: low 1-2.33, medium 2.34-3.67, and high 3.68-5. The text vs. images group consists of heavy metadata users. If a user answered 4 or 5 on questions three and four, they were ranked as heavy users of both image and text metadata. Consequently, if they answered 4 or 5 on either question three or four, they were ranked as heavy text or heavy image metadata users. If the user gave a lower score than 4 for both of the questions, the user was ranked as a non-heavy user.

The survey analysis began by calculating the mean, median, and standard deviation for each question. The answers were then analyzed in more detail by using the groups specified above. The purpose of these tests was to find significant differences between the groups discussed above. The tests were performed by using ANOVA and MANOVA analysis in SPSS software. Levene's test of equality of error variances was used for both ANOVA and MANOVA analysis. Box's test of equality of covariance matrices was also used in addition to Levene's test for MANOVA tests. If the results indicated that a significant difference was present between groups, the test was further extended by performing a post hoc test on the groups to identify where, and if, there actually was a significant difference present. Three different post hoc tests were chosen to complement each. Ryan, Einot, Gabriel and Welsh Q (REGWQ) procedure has good Type I error rate control and is usually a suggested method, but is only applicable if sample sizes are even.

Hochberg's GT2 and Gabriel's tests are more suitable for data consisting of different sample sizes. Gabriel's method was used if sample sizes varied slightly, while Hochberg's GT2 was used when bigger differences were present. If there was any doubt that population variance occurred, none of the mentioned methods could be used as the result would not have been reliable. Games-Howell's post hoc test was used in such cases.

# 5 RESULTS

## 5.1 Modeling of current editorial processes and metadata life-cycle

The complete models of the editorial processes of the five research sites can be found in Appendices 7-11. To assist the visual and textual representation of the findings, the editorial process was divided into four sub-processes: planning, production, publishing, and archiving. The following section discusses the findings within each step separately. Parts of the models are shown here, but for a complete overview of the processes the reader is advised to look at the appendices. The models are based on the data that was collected during the observations.

### 5.1.1 Planning

Planning within the editorial process is where the content creators and other involved actors, e.g. journalists, separately or together, discuss, plan, and prepare for content production. The material used during this stage is highly dependent on both the nature and type of material that is to be published. The amount of time spent on this stage varies accordingly. Ideas, repeating events, and incoming material in different forms such as user generated content may be used as a source for production. The nature of incoming material differs to a great extent. It can be sudden and unforeseen news information, which needs to be processed promptly. On the other hand, it might consist of verification for the outcome of an already known event, the result of the presidential election for example. User-generated content and feedback is another type of incoming material. This type of information needs to be assessed based on its value and reliability. Information produced by end users has enormous value in the sense that it can be instant, hands-on information, from anywhere in the world. On the other hand, how should it be used, if the content cannot be confirmed or if the reliability of the source is unknown? Especially those media companies that focus on rapid delivery of media are faced with this dilemma. The amount and type of incoming material consumed by media companies, and the pace at which it is received, is highly dependent on the nature of the media company and its products. See Figure 12 for an example. The upper process aims for prompt delivery of new information, while the lower process utilizes more in-depth planning and preparation before publishing its content. The more time there is available between publications, the more information can be actively retrieved.

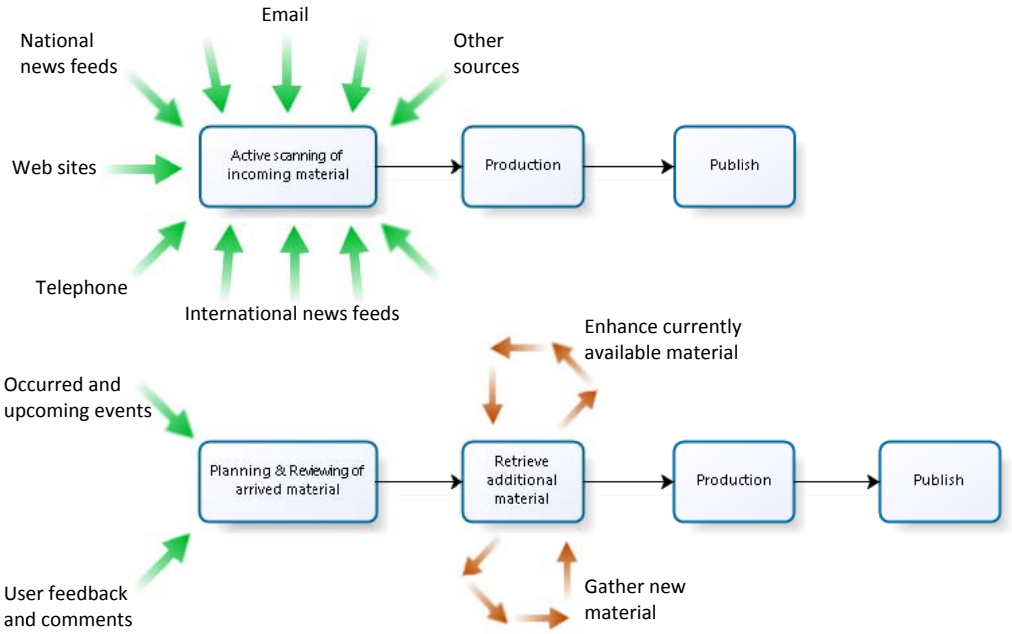



**Figure 12. Differences in obtaining initial material for content creation**

Within the five editorial offices that participated in this study there were clear differences in how they proceed towards producing content. The news agency continuously scanned incoming material from both international and national sources. This included following material published by international, well known news agencies and news web-sites. Most of this information was delivered directly into the editorial system. Other means of communication, such as e-mail and telephones were also used when receiving material. Reliability of the source was of high importance. International content arriving into the system contains attached metadata, but this information was not transferred forwards into the agency's own editorial system. Currently this information was used as read-only. The system users were not interested in using this metadata because of its varying quality and differences in the data between sources.

When comparing the model of the news agency with the other case companies, one obvious difference is found in the length of the content creation process. The news agency's editorial process was short, and most of the process consisted of the production of content. Planning played a small part of the entire process. Scanning of initial incoming material can be seen as part of planning, but actual planning took place after initial publication for creation of enhanced content versions. Planning at this stage is focused on enhancing the current content. See bottom of Figure 13. Notifications of upcoming events were also stored in advance, which can be seen as part of planning.

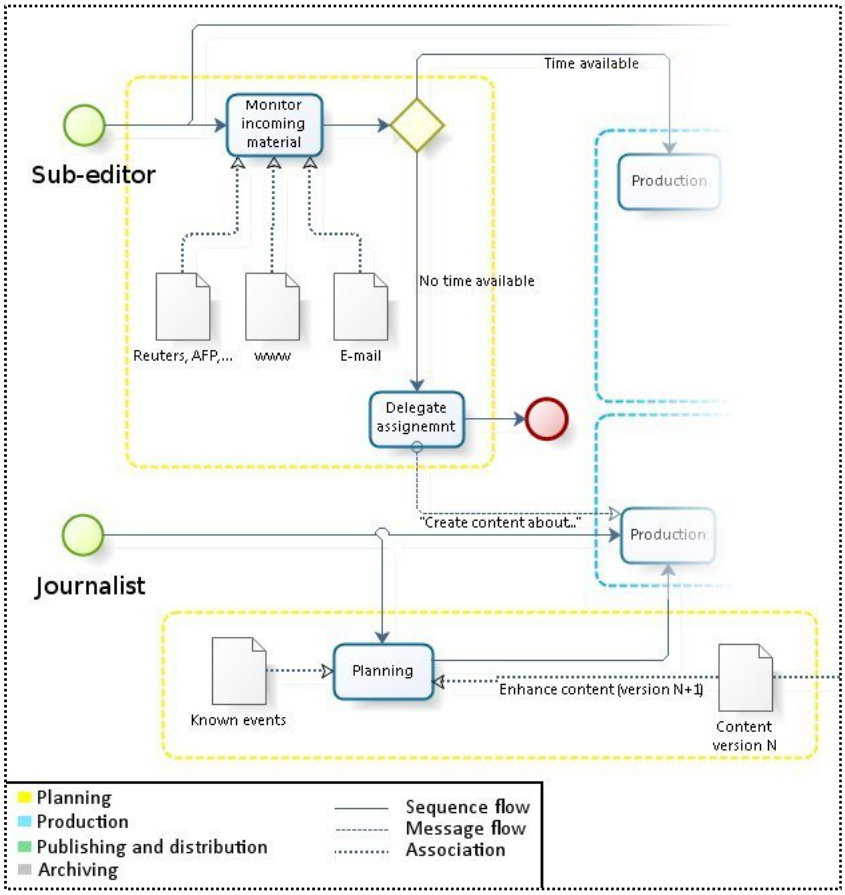



**Figure 13. Planning process in a news agency**

The published content was enhanced, expanded, and re-published as a more in-depth version later on. The reason for this was that fast delivery was set as high priority. The metadata attached to the previous version of the content was not automatically transferred to the enhanced content version. It was up to the user to manually transfer it onwards. Specific software was used for planning content, and the software included metadata fields. The metadata that was entered into the software however, was not automatically transferred to the editorial system in which the actual content then was produced.

Planning process of the weekly supplement at the newspaper was different from the news agency's planning process. The pace for consuming incoming material was not as fast since the publication time was fixed to a weekly basis. The amount of incoming material was also smaller, as most of the produced content is based on known events and other up-to-date topics. Planning took a bigger part of the production process, and involved more actors. No data was entered into the editorial system at the planning stage. See the model in Figure 14 for an overview of the process.

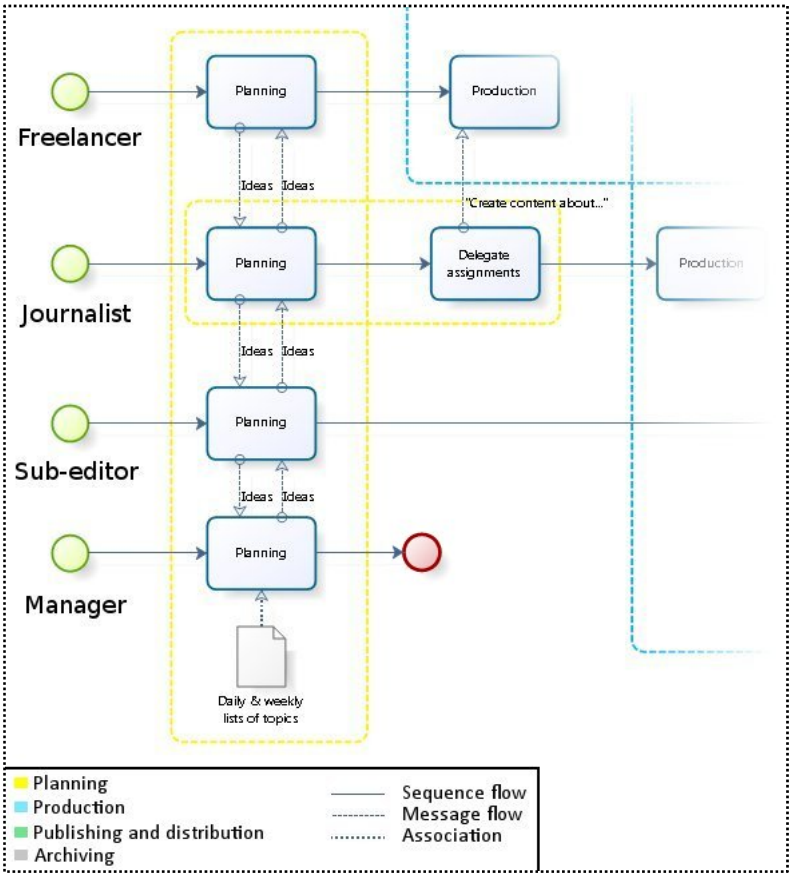



**Figure 14. Planning of weekly supplement in newspaper**

The pace of incoming material was also lower in all three magazines compared to the news agency, partly due to the fact that the publishing rate is substantially slower. Planning can be done much more in advance, and serves as a big part of the entire process. There were variations between magazines however. One of the magazines that participated in the study focused on computer and entertainment technology and products. This field and its products are continuously changing and evolving, which raises the need for more frequent planning of upcoming productions. Figure 15 visualizes the planning process of MB. The model clearly shows how there were significantly more steps involved in the planning process of this magazine, than for instance in the planning process of the news agency or the newspaper. This magazine was also heavily focused on producing cross-media content, and it was reflected in their planning process. They used to web-publish excess content that was originally written for print. Currently they were using web-publishing to support the print publication. This allowed them to offer services to the reader not possible in print. To offer such services, cross-media publishing clearly needs to be taken into account during the planning process. Magazines also tend to have sections for user input, which does not have the same requirements on reliability as when publishing news. As can be seen in Figure 15, user feedback and ideas were used in the very beginning of the planning process in the technology focused magazine.

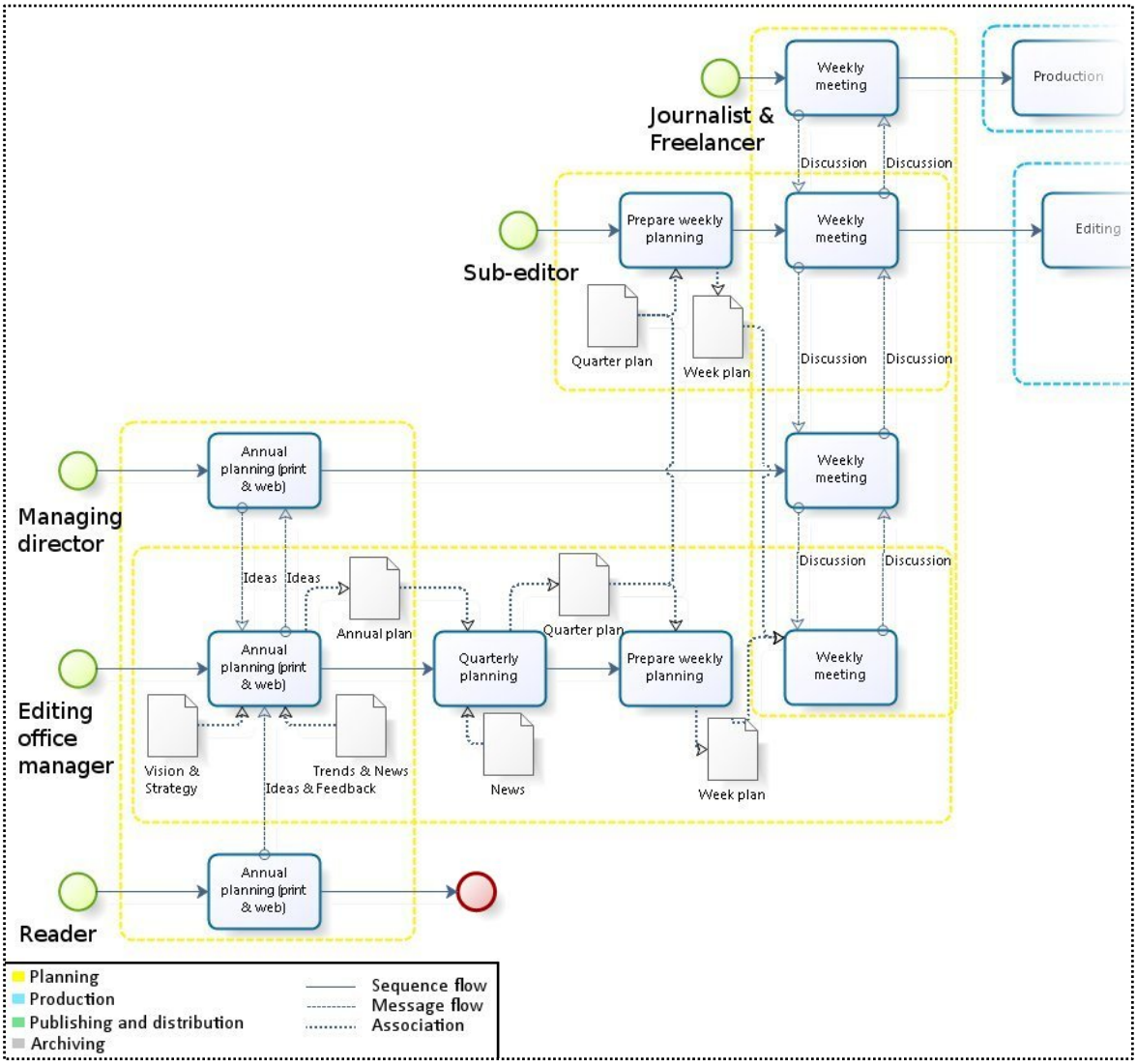



**Figure 15. Planning process of MB**

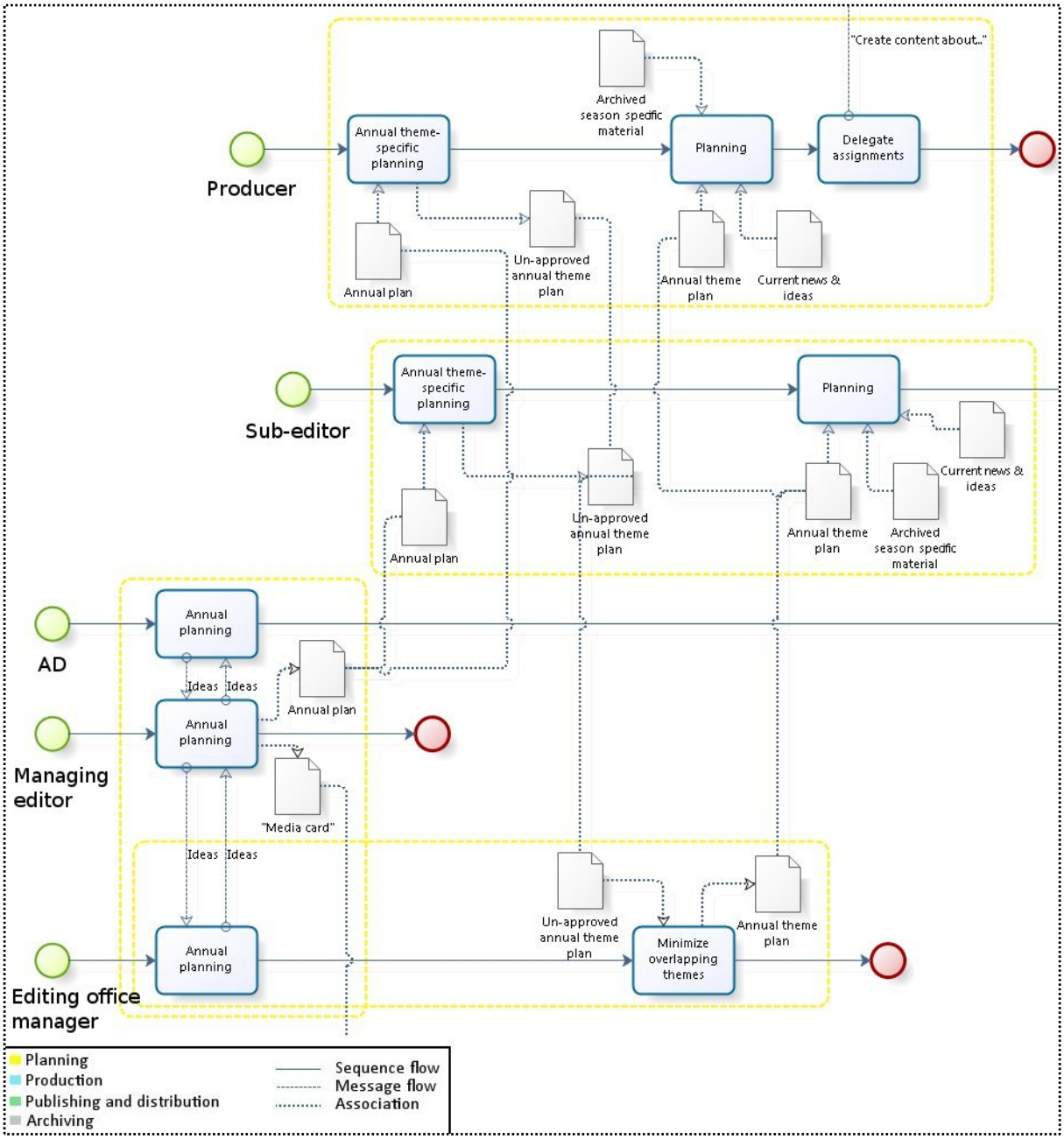



**Figure 16. Planning process of ET**

Both ET and MB magazines have similar planning processes. The planning process of ET (Figure 16) does however not take cross publication into account. One of the case magazines, Custom Publishing, focused on producing B2B content. Each product was highly customized based on the requirements set by the customer. The content produced consisted of a printed magazine or content published on the web. True cross-media publishing was not established to the same level as in the technology focused magazine discussed earlier. The most noticeable discovery within the planning process of this magazine was that the customer had a very active role during the planning process, and was actually in charge of when the production phase was allowed to begin. This was not the case in any of the other case companies. See Figure 17 for a visualization of the customer's strong role in the planning process. When the customer has accepted the brief,

the producer selects an appropriate freelancer to write the story based on the required content.

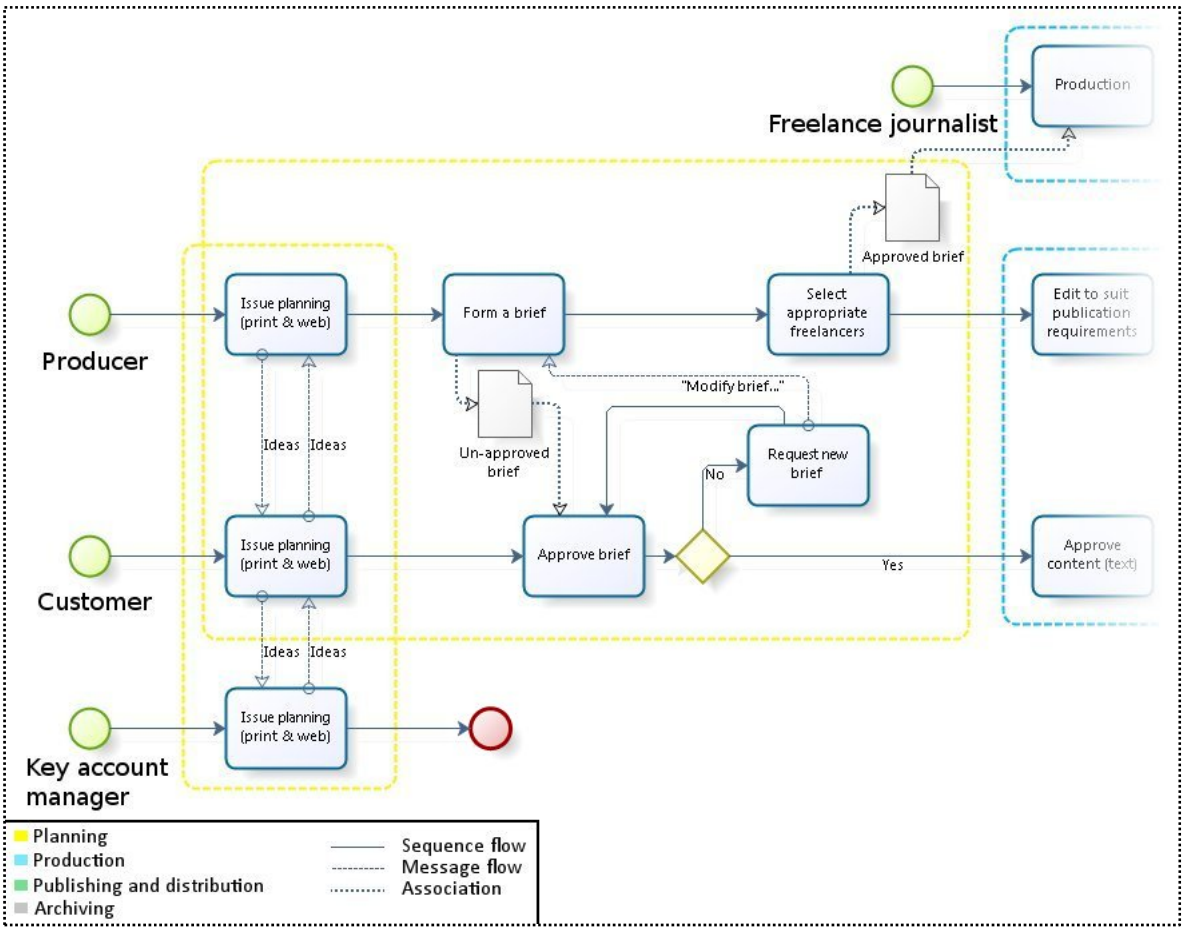



**Figure 17. B2B content planning**

Incoming text material currently has very limited amounts of metadata attached, and even if it has, it is currently not used in the planning process. None of the magazines visited currently generate any form of metadata in the editorial system at this stage.

### 5.1.2 Production

Production is where the actual content that is to be published in some form, and distributed via some channel, is generated. This process may include different types of data gathering, such as observations, interviews, and gathering of background information.

The news agency focused on rapid production and publishing of current news. The sub-editor scanned incoming material, and when something with potential value was discovered, she either began producing content based on the initial material, or forwarded the task to a journalist. There was also another journalist on call for urgent news. The first version of the content was commonly short, and editing was usually done by the author. The content was then sent to the sub-editor for final editing and proofing, who also had legal responsibility of the genuineness of published content. The metadata fields were checked at this stage by the sub-editor and lacking information was inserted. Other duties

of sub-editors varied from media company to media company, but proofing was consistently done by this person in the production process in all editorial offices.

One special practice was discovered in production process of the news agency which was not carried out in the other media companies. The content flowed iteratively through the production and publishing processes, becoming more enhanced each time. This is possible due to the fast publishing rate, as the content evolves during iterations but still is perceived as current news. This iterative process put some special needs on the editorial system, and especially metadata was not handled optimally by the system. In the same manner as during planning, metadata was not automatically transferred to the new version of the content. The responsibility of transferring this information onwards was left to the journalists/editors. The production process of STT is presented in Figure 18.

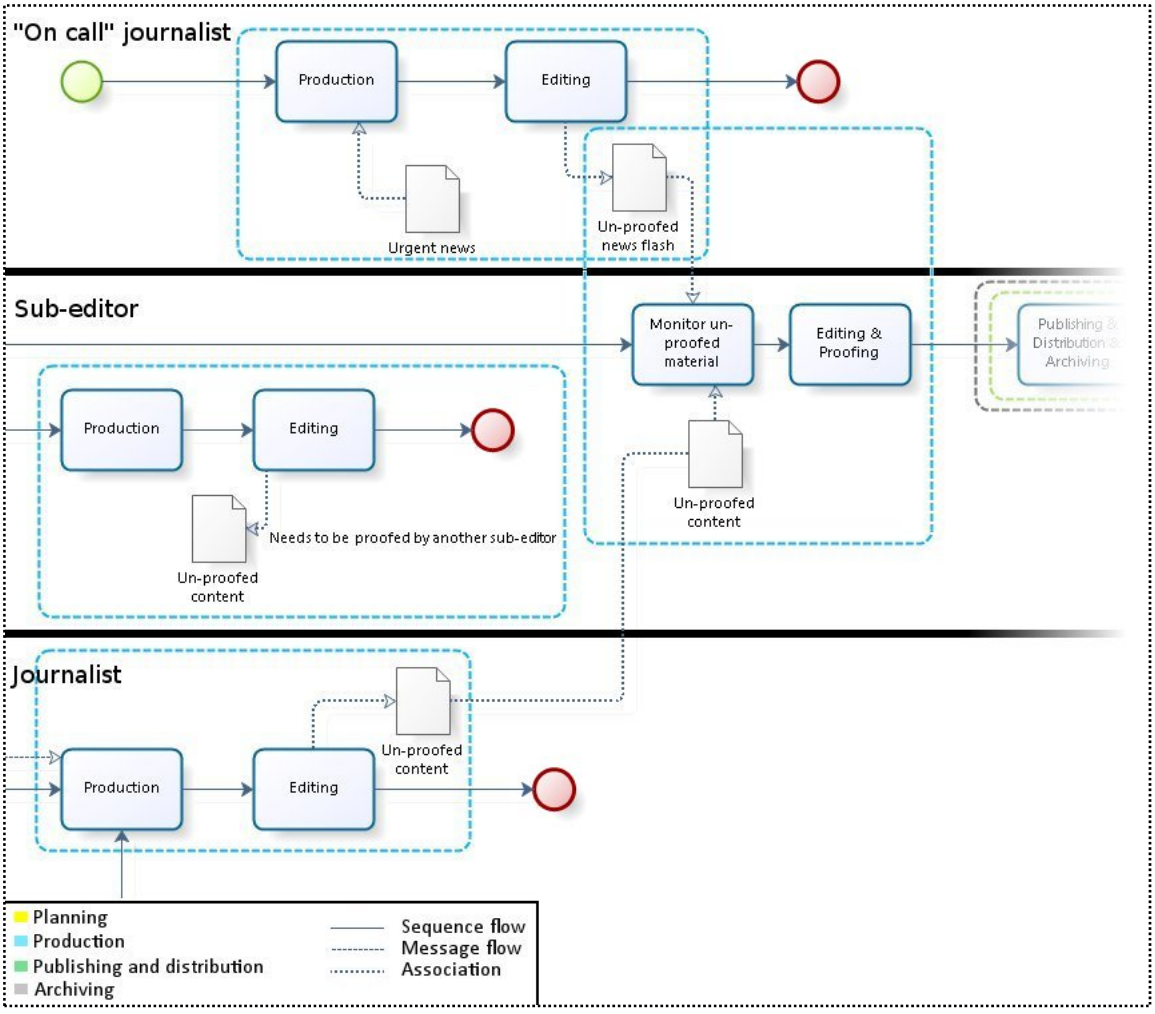



**Figure 18. Production process of STT**

The production of the weekly supplement of the case newspaper included some phases that were not present in the production process of the news agency. Similarities and contrasts between these two can be discovered by comparing Figure 18 and Figure 19. The sub-editor does not create content, her work focuses on editing and proofing of material. The journalist also has access to assistants, to whom assignments can be delegated. The journalist was also responsible for editing the assistants' material, before forwarding it to

the sub-editor. Production was both print and web-based. However, print still had higher priority. Only occasionally would the content first be published on the web. The published material was usually used as a basis for producing web-content, but identical material was never published more than once. The journalist noted in the content metadata if it had potential for being published on the web. This was then noticed in the online department, which edited the content to suit web-publishing. The same principle of first creating material for print, then for the web, was also true for the ET magazine, which can be seen in Figure 20.

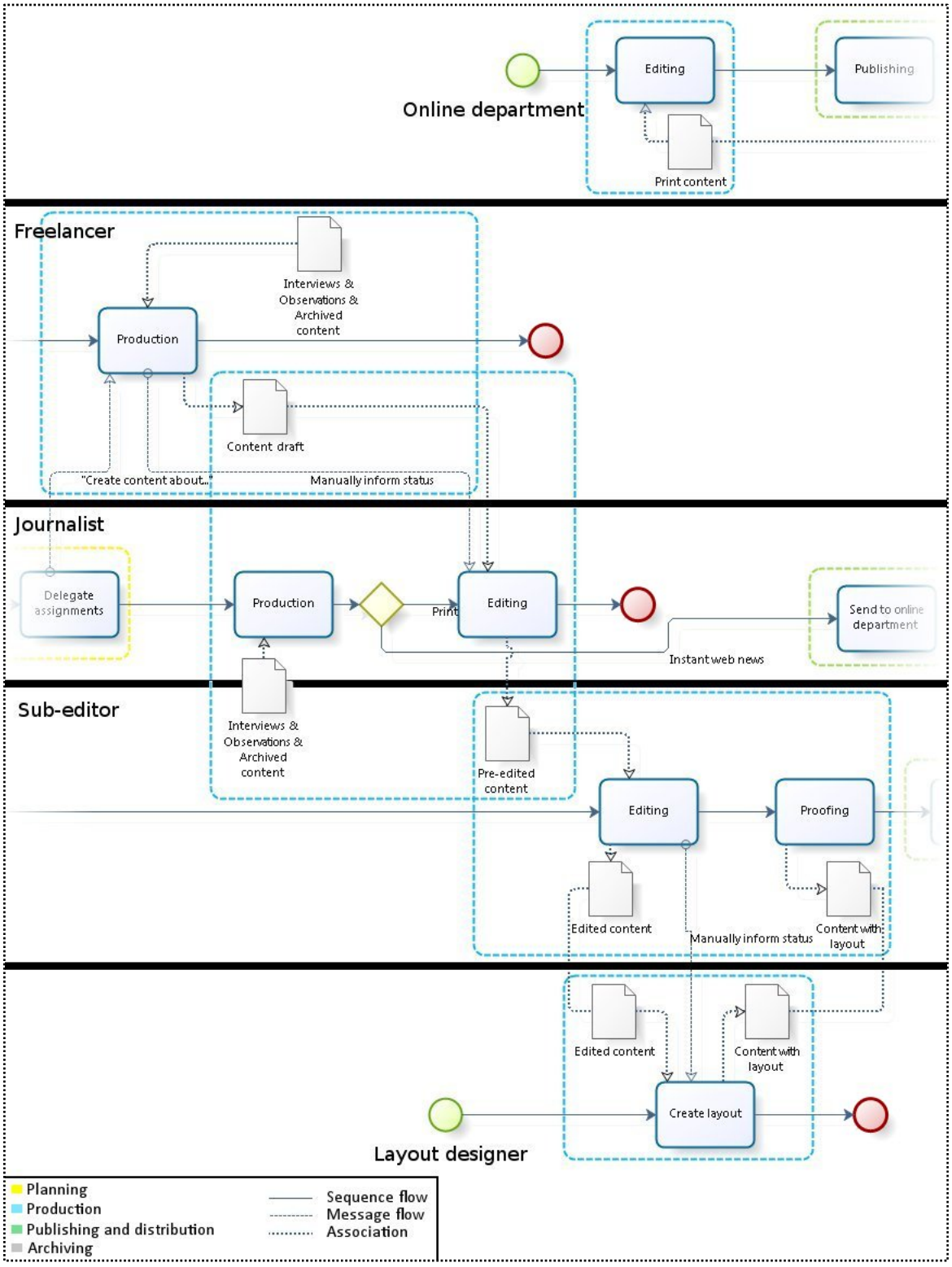



**Figure 19. Content production in a weekly newspaper supplement**

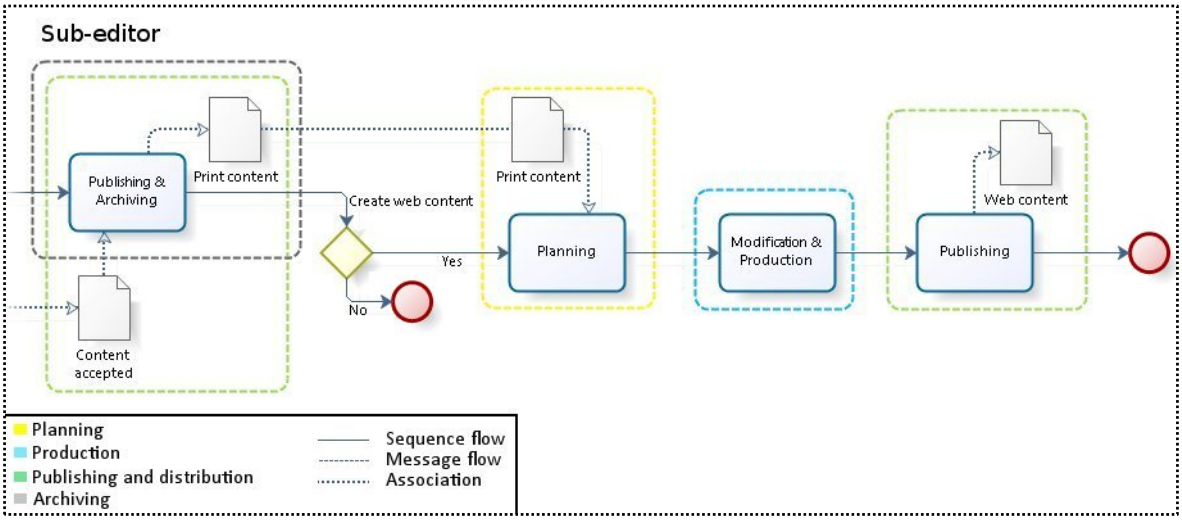



**Figure 20. Content production for web publishing in ET magazine**

The print production process of magazines ET and MB was very similar. Custom Publishing however, showed some interesting features. As mentioned in the previous section where planning was discussed, Custom Publishing produces B2B-content. During planning the customer played a strong role, and this was true also for the production process. Actual content was created by freelancers, and then sent to the producer for editing. If it did not meet the requirements set by the producer, the freelancer was requested to modify the content. Minor editing was done by the producer to fit the style of the publication. The edited content was then sent to the customer for approval. The customer either accepted or rejected the material. This process was then repeated until the customer was satisfied with the content. If the customer required multilingual content, the content could be sent for translation at this stage. Images and layout also required approval by the customer. For web-publications the graphic artist was in direct contact with the customer. For print publications the producer first edited the material when the layout had been finished by the graphic artist, and then sought approval from the customer. Due to the heavy interaction between the producer and the customer, special software was used during the production process. Both producer and customer had access to the files and could leave comments online through the software. In addition to this communication, the producer manually created a weekly status report that was sent out to the customer and key account manager. Because of the complicated and iterative communication between involved actors, and coordination of different projects, the producer relied on a spreadsheet table for keeping track of how the productions progressed. There were obvious opportunities for improvement in this system. See Figure 21 for a visualization of the production process of this magazine.

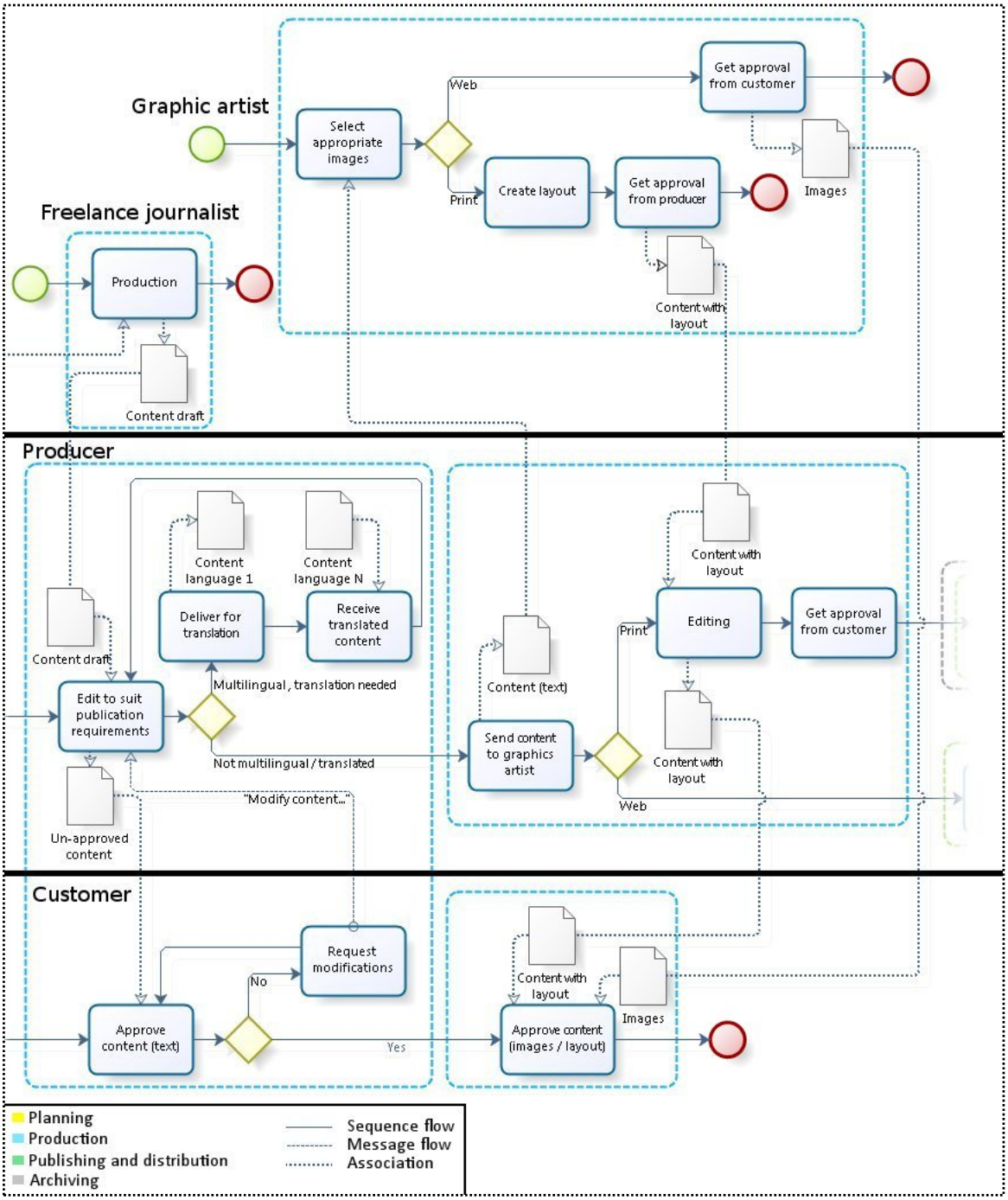



**Figure 21. Customer interaction in production process of Custom Publishing**

As mentioned earlier, ET and MB had very similar print production processes. MB however, had a webmaster for publishing and maintaining web content. The reason for this is that the magazine produces large amounts of content for the web, in addition to the magazine content. The content that was produced for web use serves different purposes than the print content. The web content could offer up to date information, interactive services such as a counter for calculating the electricity consumption of home electronics etc. The web production process was still remarkably similar to the print production process. This is because of the extensive cross-media planning that occurs in this magazine. There was no need to plan or modify content after it had been published to suit web use, since this was done before the actual production process. This practice is clearly

shown in Figure 15. Although the flow and generation of content was similar in MB and ET, MB did not use any metadata during the creation of textual content. The reason for this was that they did not employ any editorial software during this phase of the content creation process. They had the most advanced cross media planning setup, and they managed to coordinate work without any true editorial system.

### 5.1.3 Publishing and distribution

Publishing is the phase in the editorial process when the content is finished to such a degree, that it can be released for consumption by the end-user. When the content is published, it is made available for the readers. Distribution covers the delivery of the content, be it in electronic or print form. Content may therefore be published, but it does not necessary need to be distributed. For instance, content published on the home page of the content provider is not considered distributed content. The scope of this study did not allow following of the distribution of print content. Electronic publishing in different forms was examined more closely, when available.

The news agency published and distributed its material in electronic form. These actions were carried out by the sub-editor. All published material was labeled with a running number, but if content had been enhanced based on already published content, the name was usually also modified by adding a number as a prefix. This prefix grew as the content matured, and did as such represent a type of version control. This manner was however carried out as a habit, and was not systematic as differences were found between the actors in the editorial process. Content that is published on the news agency's web site was updated and overwritten by the newer versions, old versions are not stored. The sub-editor performed proofing of both content and metadata during the production of content, but vital metadata fields were scanned automatically for errors before being published.

The content was packaged and distributed to customers based on the type of service that they had purchased. News feeds with separate versions for newspaper and online publishing are typically delivered to the customer via FTP. Other forms of media, such as photos, graphics and video are also offered to the customer. ERP (Enterprise Resource Planning) information can also be purchased. Other alerting and press release services are also included in the news agency's product portfolio. Mobile news services were very popular among the customers. The news agency does not only offer alert services and distribution via different channels. Customers can also order specific content based on their needs, which is based on essence metadata. Publishing of urgent news was done instantly, but publishing could also be scheduled. Scheduling could either be done internally, or, content could be delivered to the customer, but contain restrictions on when the customer had legal rights to forward or publish the content.

The other case companies had different approaches to publishing and delivering content. The weekly supplement of the newspaper normally publishes their content in print form first. Then, the online department edited that content to suit web-publishing. In rare occasions the journalist would send a shorter version of the content directly to the online department for publishing, and then write the print version. ET published both print and

web material, but the focus was on the print content. Restructuring of the editorial process to support web-publishing at a higher level of activity was under consideration. In this magazine it was the sub-editor who was in charge of modifying print material, or creating new material, to suit web publishing. This activity within the editorial process had relatively low priority.

As in ET magazine, the sub editor of MB was responsible for publishing print content. However, the sub-editor's activity within the editorial process ended here, as there was a webmaster responsible for publishing web content. The published web-content was diverse, but possible distribution channels were not utilized to the same extent as in the news agency. In Custom Publishing the producer took care of both print and web-publishing. The difference between Custom Publishing and ET in the publishing process was that the producer of Custom Publishing did not modify the print content to suit web-publishing. The content was either produced for web use or for print, and this was decided from the beginning.

### 5.1.4 Archiving

Archiving is the process of storing produced content, and it can be performed during different phases in the editorial process. Different versions of content may be archived during the production process. Images for instance, are often stored instantly when they are inserted into the editorial system, to ensure that the original image stays intact. The archival department of the newspaper was the only archival unit included in this study. Archiving was automated to a high degree in all the case companies, in the sense that once the material was sent for publishing, it was simultaneously archived. This was true for print content, but e-published content was not always archived.

The archiving process in the news agency was entirely automated. Once content was published, it was simultaneously stored in their archive. The sub-editor was responsible for making sure that all required metadata fields were filled in before publishing. The same content was also automatically scanned for errors before publishing. By carrying out these tasks before publishing content, the amount of post-production work was reduced if not eliminated altogether. The archive process was similar in the newspaper in the sense that the material was automatically archived once it had been published. However, not all web-published material was archived, and all print material underwent manual examination after being added to the archive. Metadata fields were examined for errors and missing data, images usually required substantially more work than textual content. Most of the work required on images was related to copyright issues and essence description.

The persons who were asked about archive related matters in ET and Custom Publishing were unsure if, and for how long, the content was stored electronically. They had issues related to the retrieval of older publications, and it was often regarded easier to find older content by browsing printed publications than searching in the editorial system. MB performed archiving differently compared to the other case companies. When the print content had been sent to the printing office, the printing office extracted data from the content and sent this organized data back to the magazine. This data was then stored in a

database, which allowed for efficient and customized retrieval for web use. Content was also archived directly into the editorial system as in the other magazines, but the staff relied on the database for retrieving text based content.

### 5.1.5 Metadata life-cycle

Metadata practices within the sites were highly differentiated, and are as such presented separately from the editorial workflow models. However, similarities in specific phases of the editorial processes between the different sites are presented in Figure 22. The metadata attributes from the metadata framework were used as cornerstones for describing the types of metadata fields in the life-cycle. A more in-depth discussion of the metadata framework and the attribute types is presented in Section 5.2.

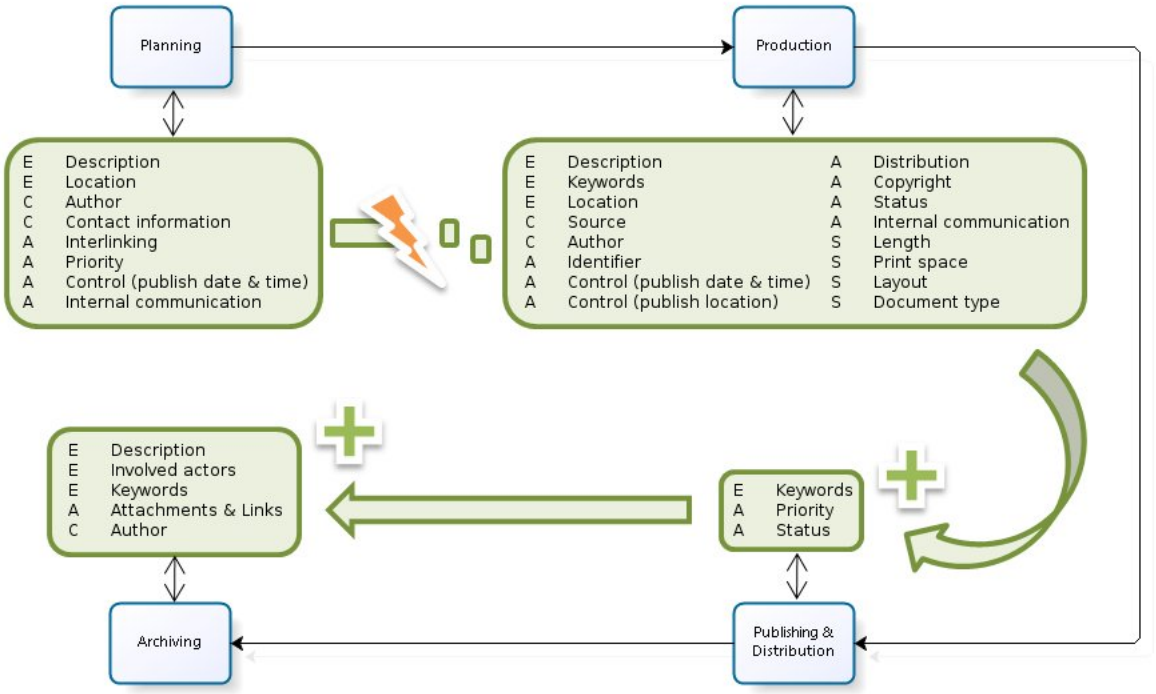



**Figure 22. Current life-cycle of metadata in editorial processes. The abbreviations in the figure stand for different metadata types: E = essence, C = contextual, A = administrative and S = structural metadata**

Figure 22 represents the current metadata creation process, and as such characterizes the possibilities of current editorial systems from a metadata creation viewpoint. The metadata field attributes show at which point they were accessible in the editorial process, but ultimately it was the user's decision if they were accessed or not. Metadata is not commonly used during planning of content, and if used, it is not efficiently transferred to the production process. The users were always offered a certain amount of metadata fields when a new document was created in or imported into the editorial system. These fields are represented by the metadata field attributes in the production phase of the above figure. Multiple metadata fields are offered at this point, but this does not necessarily mean that users insert values into in all fields at this point. Identifier, control, source, author, and status attributes were always specified at this stage. Use of essence based metadata fields at

this phase varied. Essence based information was usually handled better for image content at this stage. The "+" sign in the *publishing & distribution* and *archiving* stages indicates that field attributes are added or values are modified in addition to the already existing field attributes.

## 5.2 Metadata framework for cross media publishing

The completed metadata framework created based on qualitative data and archive samples is presented below in Table 15. More specific details on the analysis process and on how the framework was constructed are presented in Section 4.3.4. Field attributes are discussed in more detail in the following sections.

**Table 15. Metadata framework based on metadata fields in current editorial systems**

| Text | |
|---|---|
| **Metadata types** | **Field attributes** |
| Essence | Keywords |
| | Description |
| | Location |
| | Involved actors |
| Structural | Length |
| | Print space |
| | Layout |
| | Document type |
| Administrative | Identifier |
| | Copyright |
| | Internal communication |
| | Priority |
| | Status |
| | Control (publish location) |
| | Control (publish date & time) |
| | Distribution |
| | Attachments & Links |
| | History (time) |
| | History (users) |
| Contextual | Author |
| | Source |
| | Contact information |

| Images | |
|---|---|
| **Metadata types** | **Field attributes** |
| Essence | Keywords |
| | Description |
| | Location |
| | Involved actors |
| Structural | Print space |
| | Resolution |
| | Colour |
| | Document size |
| | Crop & scale |
| | Document type |
| Administrative | Identifier |
| | Copyright |
| | Internal communication |
| | Status |
| | Control (publish date & time) |
| | Control (publish location) |
| | History (time) |
| | History (user) |
| Contextual | Photographer or author |
| | Source |
| | Date & Time |
| | Photo series |
| | Location |

The metadata framework is an abstraction of the actual metadata fields found within the editorial processes of the case media companies. As can be seen from the table above, the distribution of different types of metadata was different for text and image context. Essence based metadata was however identical. Types of used fields per site and content type can be seen in Table 16 and Table 17. As mentioned earlier, MB was left out of the analysis as a traditional editorial system was not used in their editorial processes in the same manner as at the other sites.

**Table 16. Distribution of metadata types for text content**

| | Essence | Structural | Administrative | Contextual |
|---|---|---|---|---|
| **STT** | 13 % | 4 % | 75 % | 9 % |
| **HS** | 24 % | 11 % | 61 % | 4 % |
| **SM - CP** | 3 % | 15 % | 76 % | 6 % |
| **SM - ET** | | 18 % | 76 % | 5 % |
| **norm. mean** | 11 % | 12 % | 71 % | 6 % |

**Table 17. Distribution of metadata types for image content**

| | Essence | Structural | Administrative | Contextual |
|---|---|---|---|---|
| **STT** | | | | |
| **HS** | 21 % | 18 % | 36 % | 26 % |
| **SM - CP** | 11 % | 16 % | 57 % | 16 % |
| **SM - ET** | 11 % | 16 % | 58 % | 16 % |
| **norm. mean** | 15 % | 17 % | 48 % | 20 % |

### 5.2.1 Essence based field attributes

Essence metadata is highly descriptive and independent of media type. This type of metadata provides the user with information about the content essence. Preferably, the user should get such an amount of knowledge about the content essence based on this metadata that the actual content need not be examined in order to know what it is about. Haase (2004, p. 204) noted that *"…[metadata] conveys knowledge about an item without requiring examination of the item itself"*.

The essence metadata fields in current editorial systems attached to text and image content were relatively similar. The keyword fields were used to describe the content essence at different detail levels. Based on findings from archive samples, keywords and descriptive metadata for text content were used scarcely. These types of metadata were used more frequently for images. The location fields specified the location visible in a photograph, or a location that was essential for textual content. The purpose of the location field in this context was not to specify the place where the content was created. In the same manner the "involved actors" field does not specify the author of the content, but the person(s) or organization(s) that are essential for the essence of the content.

### 5.2.2 Structural field attributes

Structural metadata contains no information about the content essence, only about format and structural characteristics. For images this included information such as resolution, colour space and other color related information, file size, print size, print layout, and other information regarding structural components. Structural metadata for text content mostly consisted of information about physical print size, print layout, amount of characters, and document type.

### 5.2.3 Administrative field attributes

Administrative metadata is used for controlling and handling content in various situations. The purpose of administrative metadata varies to a great extent, as can be seen from the

field attributes for this type of metadata (Table 15). The “identifier” field consists of unique ID:s, names, and other fields that were used for identifying content. Copyright related issues were more commonly used for image content than for text content.

One important discovery made during the visits was the need for internal communication through the editorial system. Currently a great deal of production related information is exchanged outside the editorial system. If the system was used for transferring this type of information, there seemed to be no common understanding of which fields that were supposed to be used for the task. Production status of content was another type of administrative metadata that was implemented with varying levels of automation and notification methods. Some editorial offices needed to manually inform co-workers when the content was ready to move forward in the production process.

### 5.2.4 Contextual field attributes

Contextual metadata describes the environment, relationships, properties and other conditions that are not related to content essence. This includes information such as author of content, source of content, and time and location of content creation. Time and location information can be more exactly defined for photographs than for text, as photographs are created instantly while text is produced over a longer period of time.

## 5.3 Characteristics of metadata fields

### 5.3.1 Metadata fields in editorial systems versus archived metadata

Figure 23 shows the number of metadata fields that was discovered in the editorial systems of the case companies. The numbers do not necessarily represent unique metadata fields, as some of the fields appeared multiple times in different software under different names. The field names occasionally varied depending on where they appeared in the system, which further made it difficult to assess the amount of overlap in the results. Each media company is represented by two columns in the chart. The left column of each column pair represents the total number of metadata elements that was discovered within that company’s editorial system. The right column represents the number of metadata fields that was identified in the archive samples that were delivered by the media companies. If an archived metadata field could not be recognized as one of the metadata fields, and the media company did not provide an explanation for the field, it was rejected from the analysis.

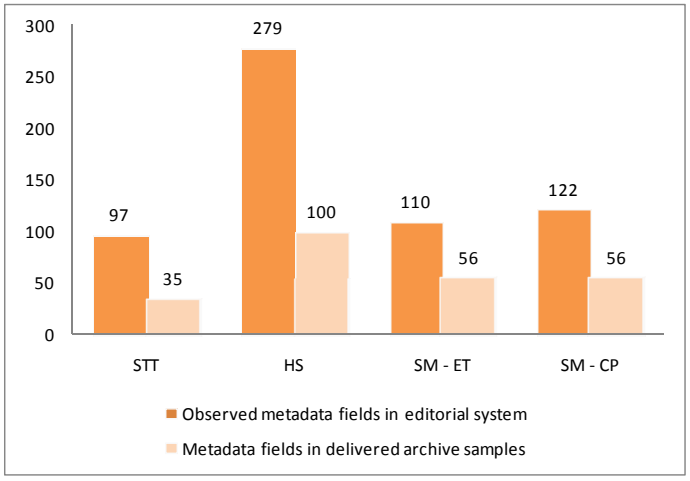



**Figure 23. Number of metadata fields in the editorial systems vs. number of metadata fields in delivered archive samples**

There were considerably fewer fields in the samples than in the systems. This is partly due to the fact that some fields were presented to the user at various phases during the editorial process and also in different software. The observed metadata field column for the newspaper includes metadata fields from the archiving department, which explains the high values. However, the newspaper also archives the biggest amount of metadata. The values for the magazines specify the amount of metadata at the very end of the production process and not the archived metadata, because most of this data is currently eliminated before the content is archived.

Because of the significant variance of discovered amounts of metadata in the media companies, the values for each company needed to be scaled in order to give equal weight to the results from the different companies. This was done by scaling the highest value from each company to one, the others values linearly according to this. The mean values presented from Section 5.3.2 and onwards are therefore normalized.

### 5.3.2 Value insertion – automation and structure

Value insertion defines how values are inserted into metadata fields. There were generally three different ways of doing this. The user was either responsible for manually inserting the value, or, the action was automated to some degree and done by the system. Two main types of automation were present in current systems. Automatic fields created values without assistance by the user. Values could be modified by the system during the editorial process, but the user was also allowed to edit the values. Fields were considered fully automatic if the user was not allowed to modify the value at any point, but these fields were regarded as a special type of automatic fields. Semi-automatic fields suggested a value to the user by automatically inserting it into the field. The user then had the possibility to edit the value, or accept it with no further action by the user. Both automatic and semi-automatic fields reduce the workload of the user. The difference between automatic and semi-automatic fields was that semi-automatic fields were meant to act as predictive fields, while automatic fields were supposed to handle information that should not require any input by the user. Table 18 displays the three main insertion ways used in

the case companies. The magazines show a high percentage of automated metadata fields, but it is worth mentioning that many of these fields are related to history information.

**Table 18. Automatization of metadata value insertion. Row sum = 100%**

| Company | Automatic | Semi-automatic | Manual | n/a |
|---|---|---|---|---|
| **STT** | 25 % | 15 % | 41 % | 19 % |
| **HS** | 18 % | 5 % | 46 % | 31 % |
| **SM - CP** | 40 % | 5 % | 24 % | 31 % |
| **SM - ET** | 52 % | 8 % | 24 % | 16 % |
| **norm. mean** | 33 % | 9 % | 34 % | 25 % |

Automatic fields tend to be related to administrative and structural metadata, as these kinds of fields are more easily automated than essence based fields. Semi-automatic fields also had a tendency to be used to a great extent with administrative metadata. However, contextual metadata fields were more prominent in this group when compared to the automatic fields. These findings might seem as insignificant, as administrative metadata covers more fields than all other types of metadata together. However, there was a clear shift in the results when examining manual metadata fields. Essence based metadata was a lot more pronounced within these fields, as can be seen by examining the results in Table 19 and Figure 24. Administrative metadata still stood as the largest group, but structural metadata was marginally manual. Interestingly, contextual metadata was inserted manually almost exactly to the same degree as it was inserted automatically. The newspaper and the news agency had implemented some degree of automatization for essence based metadata. All these fields contained location-based information. The semi-automated fields at the newspaper were linked to a tool that offered search functionality for extracting location data from text content.

**Table 19. Distribution of essence, structural, administrative, and contextual metadata according to insertion type. Row sum = 100%**

| | Company | Essence | Structural | Administrative | Contextual | n/a |
|---|---|---|---|---|---|---|
| **Automatic** | STT | 8 % | 4 % | 63 % | | 25 % |
| | HS | | 14 % | 48 % | 4 % | 34 % |
| | SM - CP | | 20 % | 63 % | 14 % | 2 % |
| | SM - ET | | 19 % | 72 % | 9 % | |
| | | | | | | |
| **Semi-automatic** | STT | | | 80 % | 7 % | 13 % |
| | HS | 20 % | | 47 % | 33 % | |
| | SM - CP | | 17 % | 67 % | | 17 % |
| | SM - ET | | 11 % | 67 % | 11 % | 11 % |
| | | | | | | |
| **Manual** | STT | 18 % | | 50 % | 8 % | 25 % |
| | HS | 22 % | 2 % | 15 % | 16 % | 46 % |
| | SM - CP | 17 % | 3 % | 34 % | 3 % | 41 % |
| | SM - ET | 15 % | 4 % | 50 % | 4 % | 27 % |

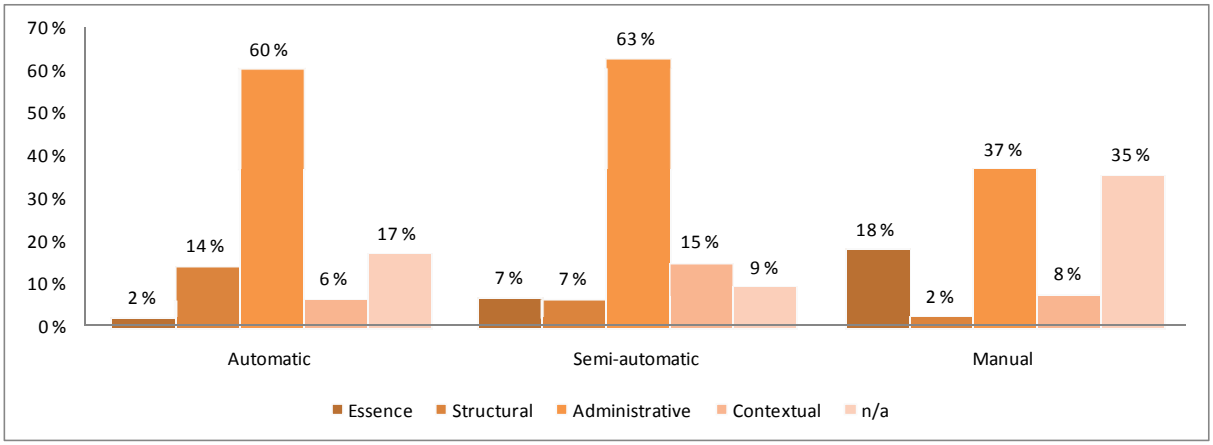



**Figure 24. Normalized mean of essence-, structural-, administrative-, and contextual metadata distribution according to insertion type**

Also form in which the values had to be inserted was analyzed. Three main ways of inserting values were discovered, but these were expanded into five groups in total. Free-form fields accepted any text input. Fixed fields only allowed insertion of predefined values, which meant that the user needed to choose the most suitable one(s) from some value pool. Fields relying on values in a hierarchical structure add some level of semantics to the entered information, but from the users standpoint fixed and hierarchical values were similarly restrictive. On the other hand, these fields produce more uniform information. Two mixed insertion types in addition to these three main types were also discovered. Certain fields offered free-form input addition to fixed or hierarchical values. Some of the fields were intentionally constructed this way, while others should have been restricted to either fixed or hierarchical values, as it caused problems if the users inserted free-from text into these fields.

Locked fields were removed from these results as values could not be inserted by users into these fields. As can be seen from Table 20, the results were similar for all case companies, with the exception of the newspaper in which a mixture of free-form and fixed fields were used to a greater extent. The magazines had a more locked fields than other case companies. Free-form text fields were most commonly used, followed by fields only allowing fixed values. Hierarchy-based fields were rarely used.

**Table 20. Different methods for inserting metadata values. Row sum = 100%**

| Company | Free-form | n/a | Fixed | Free-form / fixed | Hierarchy | Hierarchy / free-form |
|---|---|---|---|---|---|---|
| **STT** | 40 % | 20 % | 29 % | 7 % | 3 % | 1 % |
| **HS** | 48 % | 20 % | 9 % | 22 % | | 1 % |
| **SM - CP** | 41 % | 33 % | 25 % | | | |
| **SM - ET** | 46 % | 33 % | 21 % | | | |
| **norm. mean** | 44 % | 27 % | 22 % | 7 % | 1 % | |

### 5.3.3 Life-cycle of metadata

Metadata fields differed a lot in terms of dynamic nature. For instance, some values were generated once and could not be modified. Only in special occasions, for example because of errors, were they modified. These metadata fields were regarded as static. Completely static fields could be extracted from the metadata inventory by examining static fields with locked values, in other words fields that the users could not modify. Examples of such fields were unique IDs and creation dates. Alteration of these values was not allowed under any circumstances. Neutral fields allowed modification of their values, but normally editing was not necessary. Values were typically entered once into these fields, but there was no need to restrict modification of the values. This could for example include internal instructions on image modifications or working name for the content. Dynamic metadata fields on the other hand regularly changed during the editorial processes, and these fields were designed to handle multiple edits during their life-cycle. Table 21 shows the distribution of the metadata types, as well as the normalized mean of the values from all sites.

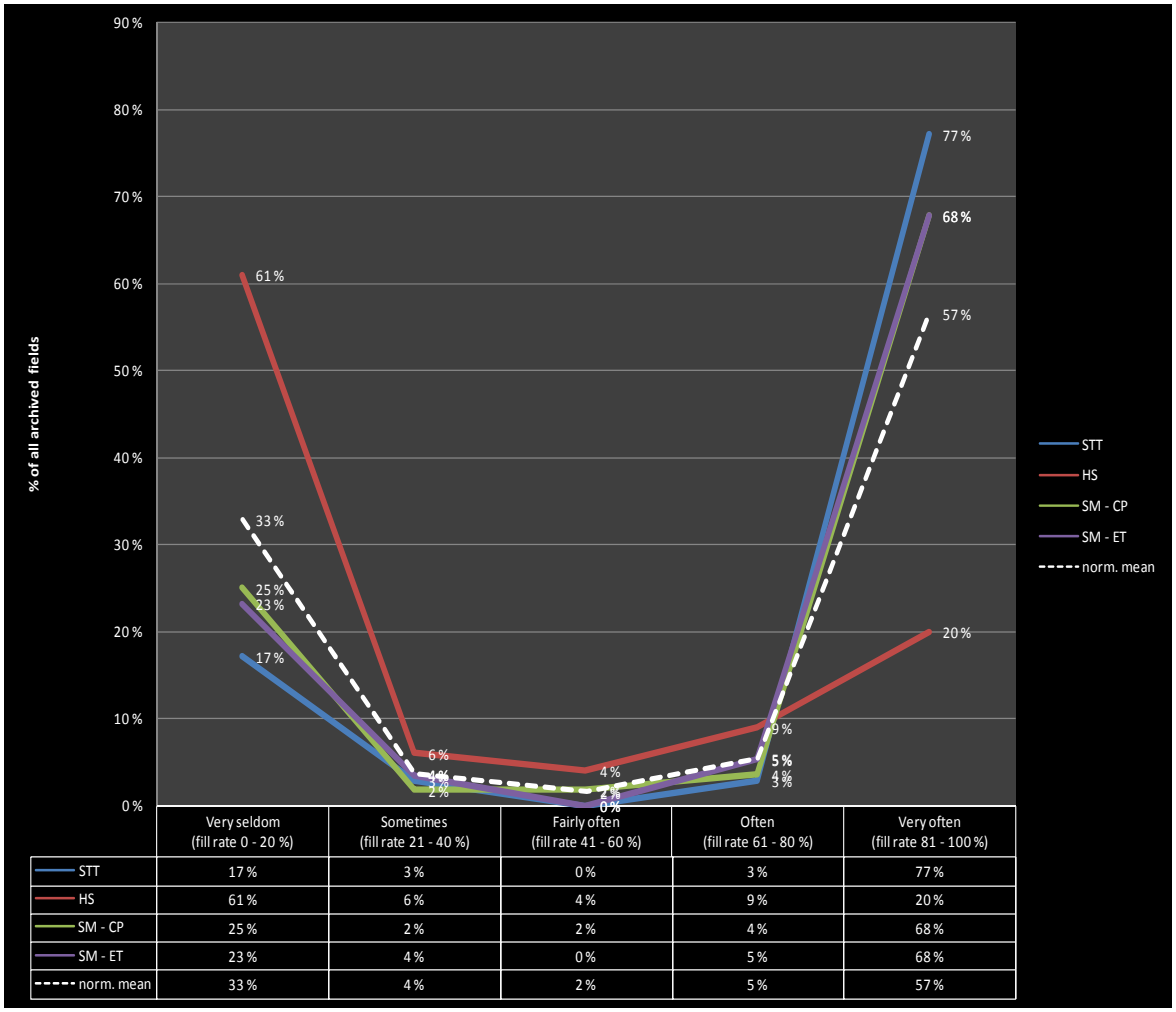
**Table 21. Nature of metadata in case companies. Row sum = 100%**

| Company | Dynamic | Neutral | Static | n/a |
|---|---|---|---|---|
| **STT** | 11 % | 38 % | 5 % | 45 % |
| **HS** | 1 % | 63 % | 2 % | 34 % |
| **SM - CP** | 20 % | 36 % | 14 % | 30 % |
| **SM - ET** | 23 % | 36 % | 14 % | 27 % |
| **norm. mean** | 16 % | 41 % | 10 % | 33 % |

## 5.4 Use of metadata fields

### 5.4.1 Fill rate of metadata fields

Figure 25 divides the fields discovered in the samples according to how often they were actually filled. For example, 17% of fields in the STT sample were only filled in 0-20% of the sample documents while 77% of fields were almost always (80-100%) filled according to the delivered metadata archive samples.




| | Very seldom (fill rate 0 - 20 %) | Sometimes (fill rate 21 - 40 %) | Fairly often (fill rate 41 - 60 %) | Often (fill rate 61 - 80 %) | Very often (fill rate 81 - 100 %) |
|---|---|---|---|---|---|
| STT | 17 % | 3 % | 0 % | 3 % | 77 % |
| HS | 61 % | 6 % | 4 % | 9 % | 20 % |
| SM - CP | 25 % | 2 % | 2 % | 4 % | 68 % |
| SM - ET | 23 % | 4 % | 0 % | 5 % | 68 % |
| norm. mean | 33 % | 4 % | 2 % | 5 % | 57 % |

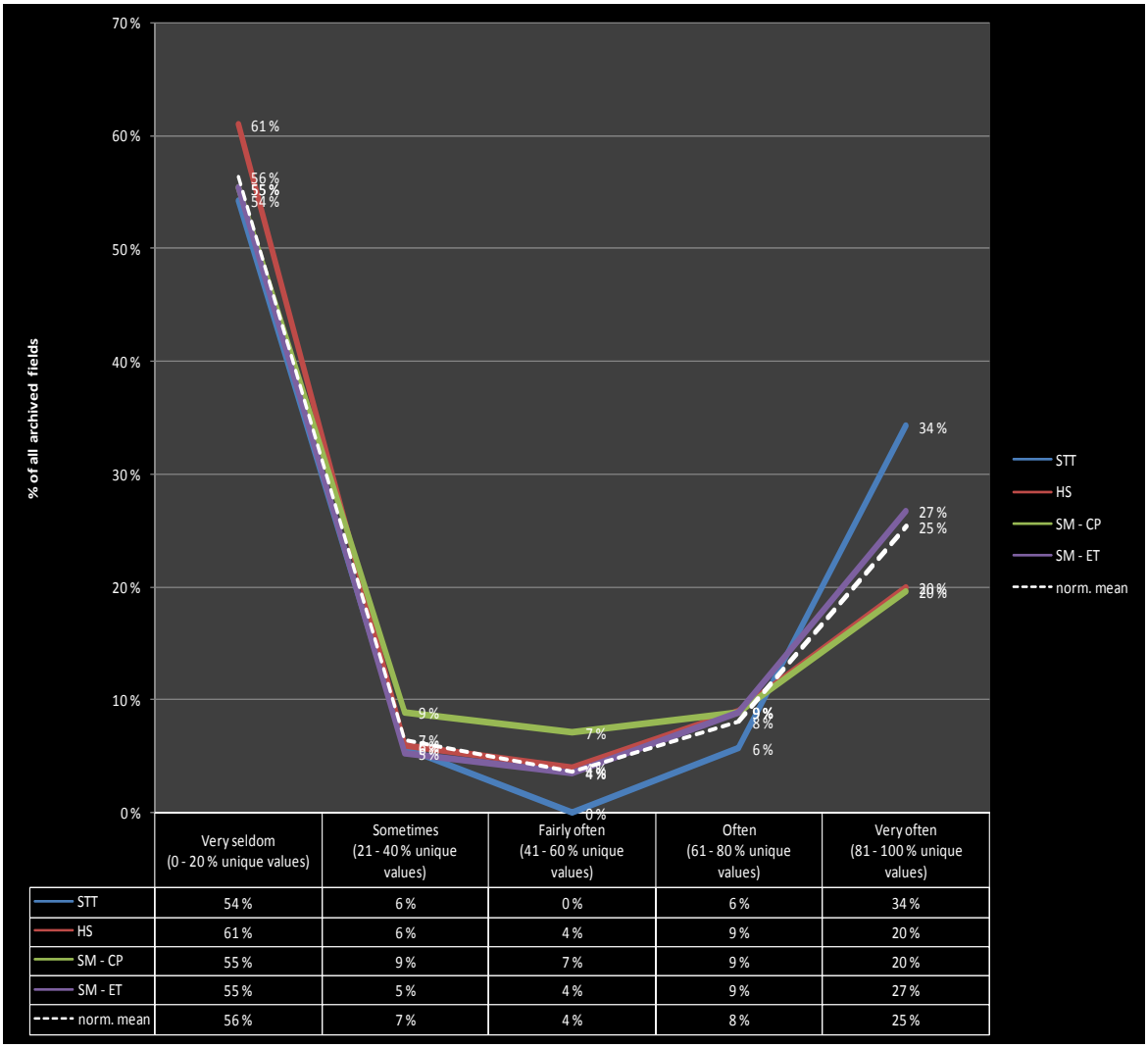
**Figure 25. Distribution of metadata fields by fill rate across the archive samples**

The newspaper had the highest amount of archived metadata (Figure 23). However, their archive samples showed the lowest overall usage. 61 percent of the newspaper's metadata fields are only used in 0-20% of all documents. The news agency and the magazines showed surprisingly similar results, even though their processes and end-purposes are very different.

The scaled mean of the metadata fill rate of the companies is also presented in Figure 25. As can be seen from the graph, the input of metadata is clearly distributed to both ends of the scale. Most fields in the archive were actively used, but 33 percent of all fields are scarcely utilized. However, studying these figures does not tell much about the quality of the inserted metadata which is discussed next.

### 5.4.2 Unique values in metadata fields

If a metadata field continuously contains the same value, the importance of that field should be questioned. The field may represent important information, but it might also be insignificant for specific types of content. The fields that showed a higher percentage of unique values were examined more closely for information about uniform use and content quality. Figure 26 shows the distribution of metadata fields in the delivered archive samples according to the percentage of unique values that they contain.




| | Very seldom (0 - 20 % unique values) | Sometimes (21 - 40 % unique values) | Fairly often (41 - 60 % unique values) | Often (61 - 80 % unique values) | Very often (81 - 100 % unique values) |
|---|---|---|---|---|---|
| STT | 54 % | 6 % | 0 % | 6 % | 34 % |
| HS | 61 % | 6 % | 4 % | 9 % | 20 % |
| SM - CP | 55 % | 9 % | 7 % | 9 % | 20 % |
| SM - ET | 55 % | 5 % | 4 % | 9 % | 27 % |
| norm. mean | 56 % | 7 % | 4 % | 8 % | 25 % |

**Figure 26. Distribution of metadata fields in delivered archive samples according to the percentage of unique values that the fields contain**

The distribution of unique values was similar in all case companies. The data was concentrated to both ends of the scale, 0-20% and 81-100%. 25 percent of all archived metadata fields showed large variations in the attached values. When these values were examined more closely, it turned out that most of the values consisted of document IDs, names and dates. 57 percent of these values were automatically inserted, and merely 16 percent were regarded as essence metadata. Most of the essence related metadata fields fall under the metadata attribute labeled “description“, in the metadata framework presented in Section 5.2. Only a fraction of the metadata fields were frequently used. This raised questions about the overall utilization of essence metadata. The usage distribution of essence related metadata fields is shown in Table 22. As can be seen, essence related metadata was modestly utilized with the majority (31% + 33%) of fields being scarcely used. 31 percent of all essence based fields in the delivered archive samples were left unfilled.

**Table 22. Employment of essence related metadata based on archive data**

| Metadata field usage % | Number of essence related metadata fields in archive | Share of all essence related metadata fields in archive |
|---|---|---|
| **0 %** | 12 | 31 % |
| **1-20 %** | 13 | 33 % |
| **21-40 %** | 3 | 8 % |
| **41-60 %** | 2 | 5 % |
| **61-80 %** | 2 | 5 % |
| **81-100 %** | 7 | 18 % |

One explanation for the large sub-section containing 0-20% unique values was that there existed many fields that were not utilized at all. Other reasons for this large sub-set include repetitive administrative metadata such as magazine or paper title and status which is fixed to "archived" at this stage. Additionally, the samples were not completely random from year 2008 as specified in the requirements. Many fields such as publishing date contained identical data, which skewed the results to some extent. The findings described above are based on combined data from the case companies, and not scaled in any way.

### 5.4.3 Comparison of objective and subjective metadata use

Users indicated the fill rate of specific fields during observations. This data was then compared to the results from the archive sample analyses. This allowed for cross-examination of utilization of specific metadata fields stated by users and actual archived metadata. Table 23 displays distribution of metadata employment according to users. There are clear differences between the companies when comparing the number of fields which according to users always are used. The scaled mean of these results is also presented in the table below. One other thing to note is that there were a large number of fields that the users were unfamiliar with. Hence the high values in the "n/a" column. The fields that were marked as occasionally or as always used by users were compared to the archived material.

**Table 23. Metadata usage according to users. Row sum = 100%**

| Company | Never | Occasionally | Always | n/a |
|---|---|---|---|---|
| **STT** | | 24 % | 20 % | 57 % |
| **HS** | 8 % | 24 % | 15 % | 53 % |
| **SM - CP** | | 6 % | 56 % | 39 % |
| **SM - ET** | 3 % | 12 % | 65 % | 20 % |
| **norm. mean** | 3 % | 17 % | 38 % | 43 % |

In total, 202 fields that were marked as "always used" by users. When compared to the archive samples, there were 109 fields that matched these fields. Of these fields 101 had a fill rate of 100%. The difference does not directly specify the degree of information loss. This is partly due to the fact that field duplicates occur in different instances of the editorial system. It was impossible to recognize all duplicates, so other means for analyzing the data were needed. It was not possible to show the exact loss of data that occurred in the

archiving process. However, a comparison of the actual fill rate and the estimated fill rate of the fields was possible. The columns of Table 24 and Table 25 indicate metadata fill rate percentage. The values in both tables represent the percentage of metadata fields that were matched from the archive samples according to estimates "always used" and "occasionally used" given by users.

**Table 24. Distribution of user indicated metadata usage ("always used"), tot. n=109. Row sum = 100%**

| Company | Fill rate 0-20% | Fill rate 21-40% | Fill rate 41-60%) | Fill rate 61-80% | Fill rate 81-99% | Fill rate 100% |
|---|---|---|---|---|---|---|
| **STT** | 8 % | | | | 8 % | 83 % |
| **HS** | | | | | 19 % | 81 % |
| **SM - YJ** | | | | | | 100 % |
| **SM - ET** | | | | | 3 % | 97 % |
| **norm. mean** | 2 % | | | | 8 % | 90 % |

**Table 25. Distribution of user indicated metadata usage ("occasionally used"), tot. n=72. Row sum = 100%**

| | Fill rate 0-20% | Fill rate 21-40% | Fill rate 41-60% | Fill rate 61-80% | Fill rate 81-100% |
|---|---|---|---|---|---|
| **STT** | 25 % | 8 % | | 8 % | 58 % |
| **HS** | 67 % | 8 % | 8 % | 6 % | 12 % |
| **SM - YJ** | | 33 % | | 67 % | |
| **SM - ET** | 20 % | 20 % | | 60 % | |
| **norm. mean** | 28 % | 17 % | 2 % | 35 % | 18 % |

The results were clear for those fields that the users ranked as always used. The fields that were archived were indeed frequently used. However, as noted before, use of essence based metadata was limited. Only three of all 109 values were essence-related. Contextual metadata was also scarce in the always used category, while the rest of the entries consisted of structural and administrative metadata. Of the always used fields 79% had automatic or semi-automatic value insertion, and none of these were essence based metadata.

The fields that the users defined as occasionally used were scattered across the scale. There was a clear explanation for the peak appearing in the 0-20% column for HS. It is generated due to the fact that many specific metadata fields existed for web publishing. This means that for each web published content there will be more empty fields than used fields.

## 5.5 Survey on metadata

The following results are based on data from the questionnaires and interviews. The sub-sections follow the structure of the questionnaire and the titles specify to which questions the results are related. The original questionnaire is found in Appendix 3. A translation of the questionnaire is provided in Appendix 4.

### 5.5.1 Use frequency of metadata based on content type (Q1-Q5)

| **Metadata in different document types** | |
|---|---|
| Q1 | I add/modify metadata daily in documents that I have created. |
| Q2 | I add/modify metadata daily in documents that I someone else has created. |
| Q3 | I add/modify metadata daily in image files. |
| Q4 | I add/modify metadata daily in articles/text files. |
| Q5 | I add/modify metadata daily in other documents. |

The first five questions dealt with metadata use frequency. Question 1 and 2 inquired about use frequency independent of content type. The results indicate that use frequency is relatively evenly spread out, but the median reveals that most of the participants actually use metadata on a daily basis (Q1, mean = 3.5, median = 5.0; Q2, mean = 3.4, median = 4.0). The reason for this difference was that participants from the magazines added less metadata than those in the news agency or newspaper. The majority of participants created metadata for text content. Results for question 5 showed a high percentage for missing and "cannot tell" answers (21%). The word "document" was perhaps not optimal in the context as the same term was used in questions 1 and 2, but with different meaning. Overall results for questions 1-5 are presented in Figure 27.

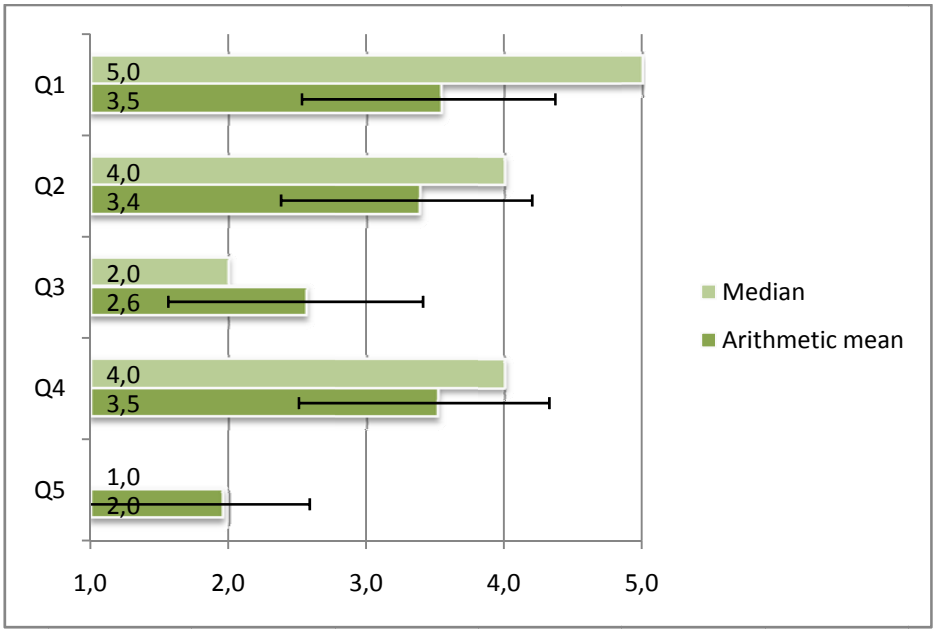



**Figure 27. Metadata use frequency based on content type**

### 5.5.2 Creation and modification of metadata (Q6-Q13)

| Metadata creation | |
|---|---|
| Q6 | I have been given sufficient instructions regarding metadata creation |
| Q7 | The reason for inserting metadata is clear to me |
| Q8 | It is clear to me who is responsible for filling in specific metadata fields |
| Q9 | It is clear to me, which specific metadata that needs to be filled in, and into which documents |
| Q10 | It is clear to me when the metadata needs to be filled in |
| Q11 | The meaning of the metadata fields are clear to me |
| Q12 | The meaning of the values available for selection in lists are clear to me |
| Q13 | It is clear to me what is allowed to insert into free-text fields |

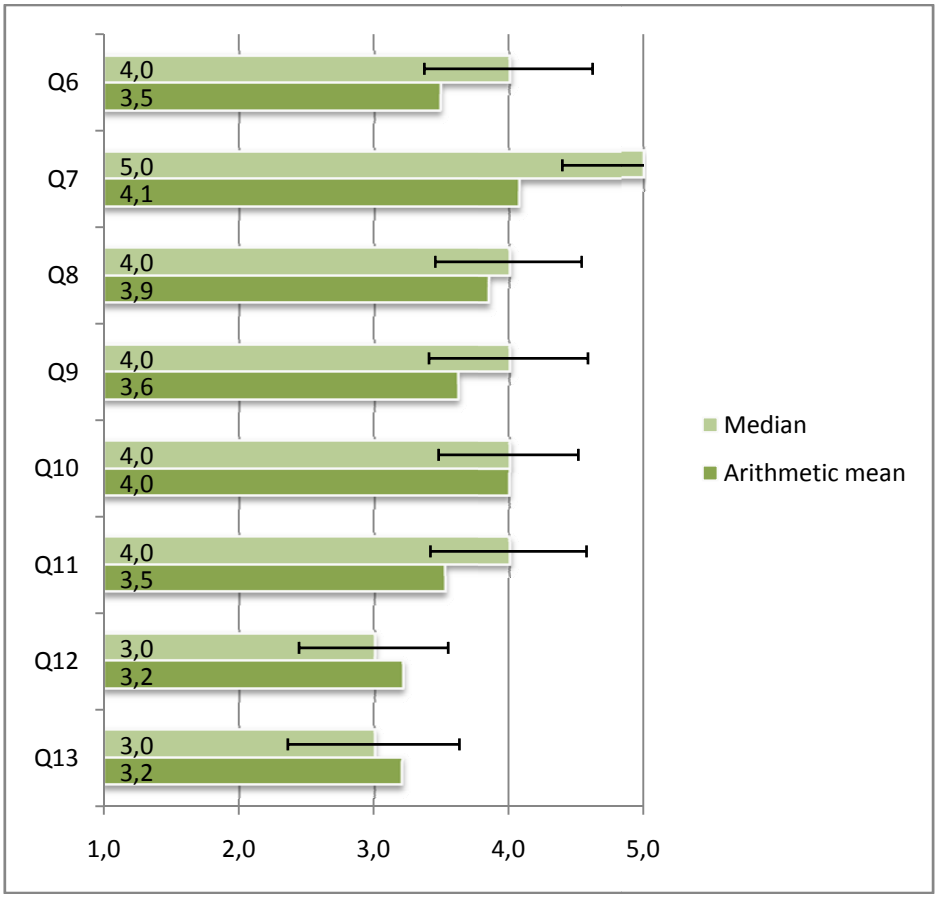



**Figure 28. Creation and modification of metadata**

Questions 6-13 focused on acquiring information about the user's knowledge of creating and modifying metadata. Overall results are seen in Figure 28. Sufficient information about how to create the metadata had been given (median = 4.0), and this result was supported by the other answers in this question group as well. The results indicated that users have a good knowledge of who is responsible for inserting the metadata (median = 4.0), what they need to insert (median = 4.0), and when it needs to be done (median = 4.0). Results also indicate that users were very well aware of why metadata is needed (median = 5.0).

Question 8 focused on knowledge of responsibilities of inserting metadata, while question 9 dealt with users' familiarity with which metadata fields that should be used for specific documents. Using Pillai's trace, there was a significant effect of metadata use frequency groups on metadata knowledge, $V = 0.13$, $F(4, 150) = 2.57$, $p < .05$. Hochberg's GT2 post-hoc test revealed that the heavy-users had significantly better understanding of responsibilities for inserting specific metadata ($p < .05$ ), as well as what kinds of metadata that needs to be inserted into specific documents ($p < .05$). However, both light- and heavy-users indicated that they had been given sufficient instructions regarding metadata creation (light-users, median = 3.0; heavy-users, median = 4.0).

During the interviews participants were asked about responsibilities for filling in metadata, and also about knowledge of currently available metadata fields. The following quote highlights the current problem of non-uniform use of metadata fields:*"if all people were like archivists and logically filled in those tickets but since we're all individuals it's a bit like shooting blanks and every now and then you find some logic behind all this."*.

Participants also answered one question about when metadata needs to be inserted or attached to content. Question 10 was analyzed with ANOVA ($F(2, 77) = 6.22$, $p < .05$), but did not meet Leneve's criteria of equality of error variances ($p < .05$). However, Games-Howell's post hoc test should be robust and was therefore performed over the data set as a curiosity. The outcome showed similar results as for questions 8 and 9. The heavy users had a significantly better understanding of timing issues than those who use metadata more seldom ($p < .05$), but this result cannot be considered reliable due to the outcome of Leneve's test.

### 5.5.3 Benefits and usefulness of metadata (Q14-Q22)

| **Benefits of metadata** | |
|---|---|
| Q14 | The majority of image metadata is useful |
| Q15 | The majority of article/text metadata is useful |
| Q16 | The majority of other document's metadata is useful |
| Q17 | I personally benefit from inserted metadata |
| Q18 | My organization benefits from the inserted metadata |
| Q19 | The readers benefit from the inserted metadata |
| Q20 | I feel that the metadata that I create is beneficial |
| Q21 | I know how the metadata that I create is used |
| Q22 | I get feedback on the metadata that I create |

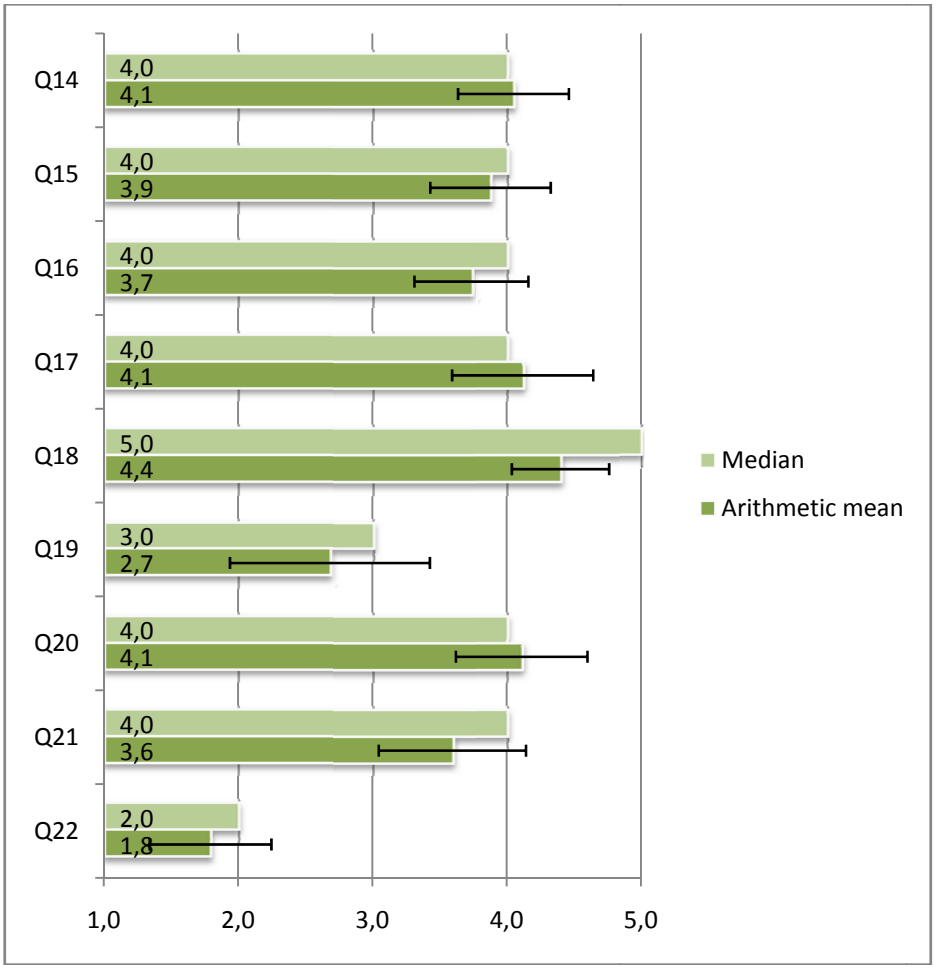



**Figure 29. Benefits and usefulness of metadata**

Questions 14-22 gathered information about the benefits and usefulness of metadata as perceived by respondents. Overall results are seen in Figure 29. Answers to question 16, which was related to the usefulness of metadata fields, contained the highest percentage of missing and "cannot tell" answers (53%). Again, the term "document" was perhaps not optimal. Overall answers indicated that metadata was useful and that participants saw metadata as beneficial for daily work both on a personal (median = 4.0) and organizational (median = 5.0) level. The answers from the interviews showed that most participants saw these tasks as part of their daily work routines. A significant effect of work group on perceived usefulness of metadata to end-users was discovered with ANOVA tests ($F(3, 55) = 4.28$, $p < .05$). Hochberg's GT2 post-hoc test showed that participants with a work position related to archiving and technical tasks (median = 5.0) indicated that metadata was more important for the reader than people working with text (median = 3.0, $p < 0.05$) and images (median = 2.0, $p < 0.05$) in the content creation process. The mean and median for people in management positions also indicated that they did not feel that the reader benefitted from the metadata (mean = 2.3, median = 1.0), but results from the ANOVA analysis did not confirm significant difference between this group and the other ones. More answers (n = 3) from people in management positions would have been needed for reliable analysis of this group of people.

### 5.5.4 Difficulties related to current metadata (Q23-32)

| Metadata difficulties | |
|---|---|
| Q23 | The same metadata is entered multiple times |
| Q24 | The same metadata needs to be entered into multiple systems |
| Q25 | Different metadata needs to be created for different customers or publishing channels |
| Q26 | The creation of metadata is sufficiently automated |
| Q27 | The creation of metadata is too time consuming compared to other work |
| Q28 | Created metadata needs to be verified/modified later on |
| Q29 | Metadata is often false or insufficient |
| Q30 | When in a hurry, I insert less metadata than usual |
| Q31 | When in a hurry, I insert less metadata than I should |
| Q32 | I enhance the metadata that I create later on |

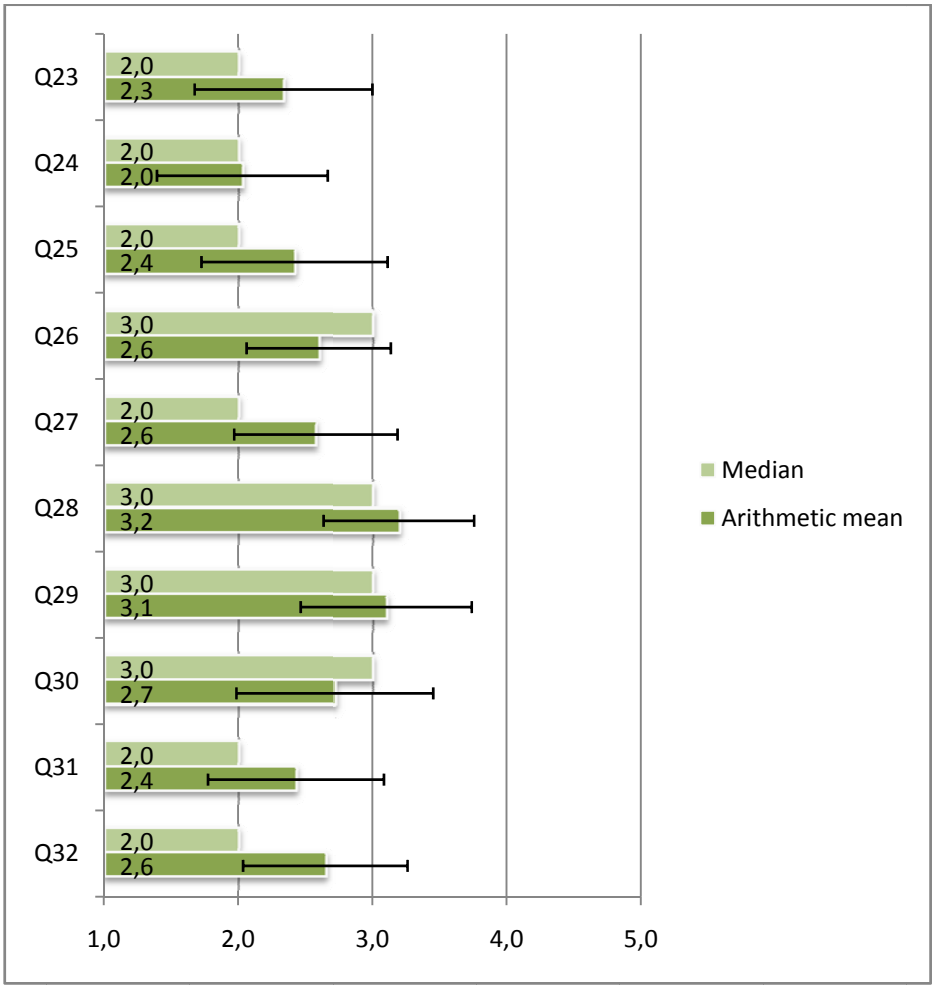



**Figure 30. Difficulties related to current metadata**

Questions 23-32 focused on current difficulties and other issues that are related to metadata. The overall results are seen in Figure 30. According to the users the same metadata does not need to be entered multiple times (median = 2.0) nor into multiple

systems (median = 2.0), and the time needed for filling in metadata is reasonable compared to other work (2.0). During the interview participants were asked about the amount of time they spend creating metadata per image / article / news item, and also if they thought that the time needed was reasonable compared to other work. The answers are in line with the results from the questionnaires. Most participants noted that the creation of metadata had become part of the daily routines. The interviewees were also asked about current problems related to metadata. The answers were a lot more negatively weighted than what the results from the questionnaires indicate. Difficulties were related to technical problems, inconsistency between different users, user interface, and insufficient automatization. Although not directly an issue, the way people communicate about production status and other content related matters were not optimally handled by current editorial systems. Interviewees indicated that they often shouted, emailed, or called others to inform about production status. Due to this the information is not stored anywhere, but people seemed unwilling to transfer this information in the system. *"Somehow I really don't feel like I need any more of those fields. Because we are so different, some use them and some don't and then the system changes. It's actually good that some of the issues are dealt with by asking other people."*

# 6 DISCUSSION

## 6.1 Goals of the study

The following sub-sections present answers to the research questions that were introduced in Section 1.2. Section 6.1.1 answers research question 1, which was related to current challenges in media companies regarding content management and metadata. Questions 2 and 4 are answered in Section 6.1.2, which covers the current metadata structure in editorial systems. Suggestions and improvements on the structure and creation of metadata are also given here. Finally, research question 3 is answered in Section 6.1.3, which discusses the current content creation processes and metadata time-cycle in the case companies.

### 6.1.1 Challenges related to content management and metadata

Current challenges regarding content management and metadata emerged during project meetings, but mainly through visits to the research sites. The amount of digital content that media companies are producing is growing at a rapid rate, and at the same time media consumption habits are evolving. The time span in which the users demand new content is to some extent dependent on content type, but generally users want to have more and more control over when they consume content. News content is created and published at short notice, and utilization of multiple distribution channels is becoming progressively important. In order to control the fast paced cross media production, new types of metadata are needed (see next section).

Consumers demand a constant stream of new content that they can access when appropriate using different platforms. This has also raised new needs for cross media publishing in magazine publishers. New metadata is not only needed for controlling the production process, advanced essence based metadata is also needed for more intelligent retrieval of information. Differences in cross media publishing practices were discovered between case companies (see Section 5.1), but all shared a common interest in improving current processes in order to better suit future cross media publishing requirements. User generated content also needs to be taken into account to greater extent in future editorial systems.

### 6.1.2 Suggestions for requirements on metadata in future cross media editorial systems

The metadata framework presented in Table 26 serves as a guideline for metadata requirements on future editorial systems. The metadata attribute types included in the framework represent current common fields discovered in the editorial processes of participating companies as well as future additions necessary for flexible cross media publishing. The bolded field attributes in Table 26 illustrate where improvements could be made on current metadata fields. The attributes with white background indicate suggestions for new metadata currently not available in the editorial systems reviewed in this study.

Keywords and descriptive metadata should be improved in order to support more advanced essence-based search and content adaptation functionalities. This is especially true for web-content, and as more and more content is produced in electronic form this is suggested as a key priority for improvement. This would enable sophisticated search and personalized publishing services among other things. The findings from the delivered archive samples showed that there were big inconsistencies in the data the users inserted into essence based metadata fields, if essence based fields were used at all. In addition to location information, essence-related timing information is suggested to be available in future editorial systems.

Contextual content related to the impact of an event is also seen as an important addition to current systems. Specifying the direction in which photographs have been taken could also be of interest. These suggestions emerged during CrossMedia project meetings.

Internal communication related to the production phase as well as production status metadata should also be enhanced. Currently editorial staff employed various channels for communicating process-related information during the production process. This information could be handled by administrative metadata within the editorial system. Future cross media publishing also creates needs for more advanced publishing and distribution metadata. Version control of content during current editorial processes was also found insufficient.

**Table 26. Suggestions for improvements (bolded) and future field attributes (white)**

| Text | |
|---|---|
| **Metadata types** | **Field attributes** |
| Essence | **Keywords** |
| | **Description** |
| | Location |
| | Date & Time |
| | Involved actors |
| Contextual | Author |
| | Impact |
| | Source |
| | Contact information |
| Structure | Length |
| | Print space |
| | Layout |
| | Document type |
| Administrative | Identifier |
| | Copyright |
| | **Internal communication** |
| | Priority |
| | **Status** |
| | Control (publish location) |
| | Control (publish date & time) |
| | **Distribution** |
| | Attachments & Links |
| | Version |
| | History (time) |
| | History (users) |

| Images | |
|---|---|
| **Metadata types** | **Field attributes** |
| Essence | **Keywords** |
| | **Description** |
| | Location |
| | Date & Time |
| | Involved actors |
| Contextual | Photographer or author |
| | Impact |
| | Source |
| | Date & Time |
| | Photo series |
| | Location |
| | Direction |
| Structure | Print space |
| | Resolution |
| | Colour |
| | Document size |
| | Crop & scale |
| | Document type |
| Administrative | Identifier |
| | Copyright |
| | **Internal communication** |
| | **Status** |
| | Control (publish date & time) |
| | Control (publish location) |
| | Version |
| | History (time) |
| | History (user) |

Interlinking of content would allow for more efficient use of metadata in future editorial systems. Essence-based search and retrieval would also benefit from this, which in turn would generate new possibilities for future products and services. Effective planning operations and content packaging allows for generation of metadata at different process phases.

Another type of interlinking was also discussed during CrossMedia project meetings. Theme based planning of content could be shared among similar content via metadata, which in turn reduces workload and allows for more detailed annotation. One possibility would be to use different abstraction levels for describing the content. The levels would be separated from each other by metadata sub-packages. Each sub-package would benefit from the already existing upper-level metadata. The metadata framework above represents metadata forming a sub-level package, while Table 27 below is an example of possible upper level metadata. For example, there could be many articles during some time span covering the same event, which happened at a specific place and time. This basic information could be shared between all related content. There should also be a possibility to integrate user generated metadata directly into the editorial system. How and to which extent this should be done is out of the scope of this study.

**Table 27. Upper-level metadata connecting multiple sub-packages**

| | |
|---|---|
| Essence | Keywords |
| | Description |
| | Location |
| | Date & Time |
| | Involved actors |
| Contextual | Impact |
| Administrative | Priority |
| | Interlinking |

The suggested metadata framework does not directly specify metadata fields. Suitable metadata elements should therefore be connected to the suggested field attributes of the framework. There exist many standards that contain applicable metadata fields, but no single standard can cope with all requirements set by current editorial work. The EXIF standard e.g. could be used for image metadata, while IPTC NewsML G2 could be suited for text content. However, even a combination of both standards they cannot provide a sufficient metadata structure for the needs of media companies. The fields currently missing from these standards are mostly related to administrative metadata, which is connected to the content creation process. In order to create a metadata structure that supports intellectual interchangeable content and is suitable for multiple media types, a mapping scheme to standards based on the framework is suggested. The framework should first be further enhanced with more specific metadata fields, creating the grammar on which the mapping would rely. This would ensure that as much metadata fields as possible are interchangeable with other standards, while offering specific metadata fields required by future cross media publishing media companies. See Figure 31 for a visualization of interaction between the framework, organizations, and standards.

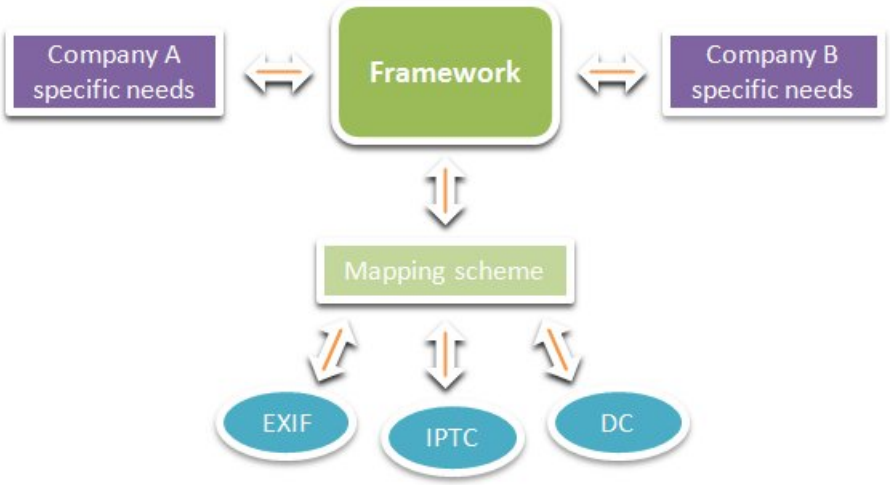



**Figure 31. Visualization of implementation of metadata framework**

Even though the fields follow standards actions need to be taken to ensure that the values can be intelligently processed. It is clear that many fields require input in free-text form, but if possible, implementation of vocabularies, taxonomies, or ontologies would provide greater possibilities for enhanced processing of content. It is also obvious that the same metadata field could be used with very different values depending on the intended purpose. Using publicly available vocabularies, taxonomies, or value schemes allows for easier exchange of content even if combined vocabularies or taxonomies suited for multiple media companies could not be formed. Creation of a mapping scheme between standards and additionally required fields, as well as exploring possibilities of shared and controlled values is suggested as a future research topic.

Based on the findings in this study it became clear that essence based metadata was not implemented to a great extent. Fields containing essence based metadata were scarce and not employed uniformly by users. Essence metadata was to some extent more implemented for images. As the amount of content is ever growing, it is becoming more and more difficult to retrieve relevant data. This is why essence based metadata, structured by a grammar is very important for text content as well. Essence metadata is probably the most time consuming type of metadata to produce, and it is very difficult to automatically extract descriptive information from content.

When designing a metadata structure covering the entire cross-media publishing process, it is obvious that there will be metadata elements that are very specific and only concern certain parts of the editorial processes. Many of the current fields were unused at least according to the participants during the observations: *”there is quite a lot which is not used at all anymore...”*. Personalized views might encourage people to create metadata if the possibilities are not overwhelming.

### 6.1.3 Editorial processes and the metadata time-cycle

Content creation is a complex process including many actors and phases. Metadata was part of the production, publishing & distribution, and archiving process, but currently scarcely used during planning of content. The time available and the length of the planning

processes varied between the case companies, which presents needs for different types of metadata. As an example, metadata could be used to keep track of the distribution of content themes for lower paced planning, which could be especially useful for magazines. In fast paced planning metadata could e.g. automatically be used to inform advertisers of upcoming content or keep track of available journalists for upcoming urgent tasks. The production process is the most complex sub-process within the editorial process, if taking into account the length of the content flow, iterations of content flow, communication between actors, the amount of involved actors, and timing issues. Significant differences between the case companies and their production practices were discovered.

Cross-media publishing was taken into consideration at very different stages in the editorial process. Some of the case companies plan their content according to the distribution channel, while others focus on creating the print version of the content first, and then proceed with modifying or creating appropriate content for other distribution channels. Most of the metadata that flows within editorial systems was generated in this sub-process, but again, some exceptions were discovered.

Contrary to incoming text, images had varying amounts of metadata attached when they were inserted into the editorial system. Metadata such as content keywords, photographer, date and time, and copyright information were sometimes attached to the content before it arrived at the media company. The most common practice for creating initial metadata and attaching it to text material was that a form was presented to the user when a new document or object was created in the editorial system. Four of the media companies were using the same editorial system at some point in the production phase. The software was however customized to suit the different needs of the media companies, which is why some have more similar metadata structures than others.

People who worked with producing text and image content generated different types of metadata during their work routines. People dealing with archiving issues were also part of the annotation chain, but people in management positions did not create or modify metadata. Administrative metadata was mostly automated, but especially essence-related metadata needed manual input by users. Most metadata fields were presented to the users at the beginning of the production process, but annotation responsibilities were left to the users as only few fields required values at this point. Most metadata fields could usually be accessed and modified at any point during the production process, which have led to diversified metadata practices.

## 6.2 Reliability analysis

### 6.2.1 Data

The qualitative data of this study consisted of observations, interviews, and questionnaires. All discussions during observations were recorded, and in case of uncertainty during analysis these audio files were reviewed. Screenshots of metadata fields occurring in the editorial systems were also taken in order to provide verification of collected data. The interviews were recorded as well, and these audio files were also transcribed.

The data obtained from the questionnaires was first examined for inconsistencies or errors before proceeding with the actual analysis. Questions with high number of missing answers or "cannot tell" answers were also examined to see what could have caused these peaks. The term "document" was perhaps not optimal within the context.

Process models were created based on observations, and are the results of an iterative process. The metadata life-cycle was modeled separately from the editorial processes, as the metadata modeling proved more difficult because of varied metadata practices. In order to create a more detailed metadata life-cycle model, additional participants would have been needed for observations.

Quantitative data consisted of archive samples that the participating media companies delivered. The requirements for the archive samples were sent to the companies in advance. The specification regarding the time span for the samples was stated in the requirements as random from year 2008. However, most of the delivered samples were from the same day or from a span of a few days. By taking random samples from a bigger time span the diversity of the samples would have been grater, offering a more realistic view of current metadata practices. Apart from this issue all archive samples were accepted as representative and genuine.

### 6.2.2 Results

The results of this study were regarded as reliable in the sense that they represent the current metadata practices from the roles represented by people that participated in this study. However, some of the data that was collected could most likely have been enhanced by adding additional participants with different work positions to the study. This would have minimized the "n/a" numbers in the results. However, this indicates that numerous fields were irrelevant and could have been hidden from the users that participated in this study. Lack of instructions about the metadata fields could also have been a reason for this.

Another issue that affected the results was the fact that the same fields occurred multiple times within the same system in different views and even different software. Occasionally these were linked to each other, but at times links between the fields were created at specific phases in the editorial process. This raised issues regarding the treatment of these types of fields. In order to treat all fields equally, every discovered metadata field was documented as a unique field. This might have led to duplicates.

Models of the editorial processes were produced based on observations. They were based on the views of the participants, and more advanced models could have been created by conducting further observations and including participants with different work positions. The models were however sent to the companies for approval, and only small modifications were made based on the comments. The results are therefore regarded as valid and reliable.

The results from the interviews were used to complement the results from the observations and the questionnaires. Not all themes were covered with all participants. A more

structured interview method might have created more data. The results that were produced were however considered reliable, and they fulfilled their task of complementing the other results.

The results from the questionnaires were also considered reliable, but conflicts between some of the answers and the findings from the observation occurred. For example, the results of question 11 (The meaning of the metadata fields are clear to me) did not correlate well with the results from the observations. According to the observations there existed a large amount of metadata fields whose intended meaning was unclear to the users. It is possible that the users answered question 11 based on the fields that they actually use, of which they had a good understanding. The same theme was dealt with during the interviews. One answer was: *"Those that we don't need to use are unclear to me, and I cannot exactly specify why we don't use them because it hasn't.. we've been told and instructed about which fields we need to insert information into, and I know why but for the other fields I could not say."*. This same person answered the questionnaire after the interview, and gave question 11 a score of 4. This would imply that at least this specific participant answered the question based on metadata fields that were used during work, and not on available metadata fields. This would explain the differences between the results.

# 7 CONCLUSIONS

In this study the current metadata practices related to cross media publishing in five editorial offices were examined. Both quantitative and qualitative data were collected and analyzed. Current editorial systems in media companies offer fairly large amounts of metadata possibilities to the users. It became clear that a large amount of current metadata fields were left untouched, either because of uncertainty of what the fields were meant for, or, because the user did not see any benefit of using them. Several of the automatically, or semi-automatically system controlled fields, could have been hidden from the user as they were never modified or used by the average user. Essence based metadata was scarcely implemented, especially for text content.

A metadata framework consisting of essence, contextual, structural, and administrative metadata field attributes was created based on discovered metadata in the case companies. Based on the results suggestions for requirements on metadata in editorial systems were given. These concern essence based date and time information, version control, direction in which photos are taken, and impact of events. Suggestions for improvements on current metadata were also given. Especially internal communication could be enhanced.

There were differences in cross media publishing practices between the case companies, but a common goal was to enhance cross media publishing. The focus was especially on improving web-based publishing in order to complement the print version of the content. In other words, identical content should not be published via various channels but the different media should complement each other and as a result improve the end-user experience. Improvements on current metadata fields are needed in order to achieve this. Another future requirement on editorial systems is the inclusion of theme or topic based planning of content by including different layers of metadata. Metadata will have an important role in reducing the needed amount of overlapping work, and will also improve services based on essence and contextual information.

The research questions were answered and the results were considered reliable. Current challenges that media companies face with regard to content management and metadata are presented. Types of metadata that are used, as well as the actors involved and the content creation workflow are also discussed. Suggestions for improvements are given based on the findings.

# REFERENCES


Alanko, T. (2009), *Metadata in Magazine Production*, Master's Thesis, Helsinki University of Technology, Department of Media Technology, Espoo, Finland. 91 p.

Anon (a), International Press Telecommunications Council (IPTC), (online), [Referenced on 2.7.2009], available in WWW-format:
<URL: http://www.iptc.org/cms/site/index.html?channel=CH0088>

Anon (b), XMP, *Adobe Extensible Metadata Platform in Newspaper and News Magazines*, (online), [Referenced on 3.7.2009], available in PDF-format:
<URL: http://www.adobe.com/products/xmp/pdfs/xmp_news-mags_wp.pdf>

Anon (c), DIG35 Initiative Group, *Metadata Standards: A Smarter Way to Look at Digital Images*, (online), [Referenced on 3.7.2009], available in WWW-format:
<URL: http://www.i3a.org/technologies/metadata/>

Anon (2002) JEITA CP-3451 Exchangable image file format for digital still cameras: Exif version 2.2, (online), [Referenced on 10.06.2009], available in PDF-format:
<URL: http://www.exif.org/Exif2-2.PDF>

Anon (2008a) CrossMedia Project Plan v1.1, Flexible Services Programme 2008 in TIVIT.

Anon (2008b) Dublin Core Metadata Element Set, Version 1.1, (online), [Referenced on 1.7.2009], available in WWW-format:
<URL: http://dublincore.org/documents/dces/>

Anon (2008c) IPTC Standard, Photo Metadata 2008, IPTC Core Specification Version 1.1, IPTC Extension Specification Version 1.0, Document Revision 2, (online), [Referenced on 3.7.2009], available in PDF-format:
<URL: http://www.iptc.org/std/photometadata/2008/specification/IPTC-PhotoMetadata-2008.pdf>

Anon (2009a), RASKE & RASKE2, (online), [Referenced on 16.1.2009], available in WWW-format:
<URL: http://www.it.jyu.fi/raske/index_en.html>

Anon (2009b), International Press Telecommunications Council (IPTC), (online), [Referenced on 17.2.2009], available in WWW-format:
<URL:
http://www.iptc.org/cms/site/index.html;jsessionid=aFyJzJo_7O54?channel=CH0086>

Anon (2009c), Specifications for Business Process Modeling Notation (BPMN), Version 1.2, (online), [Referenced on 28.5.2009], available in WWW-format:
<URL: http://www.omg.org/docs/formal/09-01-03.pdf>

Besser H., Van Malssen K., *Pushing Metadata Upstream into the Content Production Process: Preliminary Studies of Public Television*, New York University, New York.

Bødker, S. & Petersen, A. B. (2007), *Seeds of Cross-Media Production,* Computer Supported Cooperative Work 16, 6. p. 539-566.

Boll, S., Klas, W. & Sheth, A. (1998), *Overview on Using Metadata to Manage Multimedia Data*. In: Klas, W. & Sheth, A., *Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media*. New York, USA, McGraw-Hill. p. 1-24.

Chakravarty, A., Cirevegna, F. & Lanfranchi, V. (2004), *Cross-media document annotation and enrichment,* SAAW2006 - 1st Semantic Authoring and Annotation Workshop, 6.11.2006 Athens, GA, USA, Web Content Mining with Human Language Technologies.

Creswell, J. W. (2003), *Research Design - Qualitative, Quantitative, and Mixed Methods Approaches,* 2nd edition, USA, Sage Publications, Inc., 246 p.

Curtis, K., Foster, P. W. & Stentiford, F. (1999), *Metadata – The Key to Content Management Services,* IEEE Meta-data 99 Conference, 6-7.4.1999 Maryland, USA.

Dailey, L., Demo, L., Spillman, M. (2003), *The Convergence Continuum: A Model for Studying Collaboration Between Media Newsrooms*, Association for Education in Journalism and Mass Communication, Kansas City, Missouri, USA, 28 p.

De Stutter, R., Notebaert, S. & Van de Walle, R. (2006), *Evaluation of Metadata Standards in the Context of Digital Audio-Visual Libraries,* 10th ECDL - EUROPEAN CONFERENCE ON RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES, 17-22.9.2006, Alicante, Spain.

Dimitrova N., (2004), *Is it Time for a Moratorium on Metadata?,* IEEE Multimedia 11, 4. p. 10-17.

Duval, E. (2001), *Metadata Standards: What, Who & Why,* Journal of Universal Computer Science 7, 7. p. 591-601.

Gilliland-Swetland, A. J. (2000), *Setting the Stage,* (online), [Referenced on 04.02.2009], available in PDF-format:
<URL: http://www.slis.kent.edu/~mzeng/metadata/Gilland.pdf>

Haase, K. (2004), *Context for semantic metadata*, In Proceedings of the 12th Annual ACM international Conference on Multimedia, 10 – 16.10.2004, New York, USA.

Hillmann, D. (2005), *Using Dublin Core*, (online), [Referenced on 10.6.2009], available in WWW-format: <URL: http://dublincore.org/documents/usageguide/>

Jokela, S. (2001), *Metadata Enhanced Content Management in Media Companies,* Master's Thesis, Helsinki University of Technology, Department of Computer Science and Engineering, Espoo, Finland. 85 p.

Kosch, H., Böszörményi, L., Döllerm M., Libsie, M., Schojer, P. & Kofler, A. (2005), *The Life Cycle of Multimedia Metadata,* IEEE MultiMedia 12, 1. p. 80-86.

Lyytikäinen V. (2004), Contextual and Structural Metadata in Enterprise Document Management, Jyväskylä, Finland, University Library of Jyväskylä, JYVÄSKYLÄ STUDIES IN COMPUTING 37, 60 p.

Mathes, A. (2004), *Folksonomies – Cooperative Classification and Communication Through Shared Metadata* (online), [Referenced on 07.01.2009], available in WWW-format:
<URL:http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>

Mauthe, A. & Thomas, P. (2004), *Professional Content Management Systems – Handling Digital Media Assets,* England, John Wiley & Sons Ltd., 314 p.

Möller, K. S. (2001), *Information categories and editorial processes in multiple channel publishing,* Royal Institute of Technology, Division of media technology and graphics arts, Stockholm, Sweden. 180 p.

Pereira, F., Vetro, A. & Sikora, T. (2008), *Multimedia Retrieval and Delivery: Essential Metadata Challenges and Standards,* Proceedings of the IEEE 96, 4. p. 721-744.

Salminen, A., Lehtovaara, M. & Kauppinen, K. (1996), *Standardization of Digital Legislative Documents,* Proceedings of the 29th Hawaii International Conference on System Sciences (HICSS) 5, p. 72-81.

Salminen, A., Kauppinen, K. & Lehtovaara, M. (1997), *Towards a Methodology for Document Analysis,* Journal of the American Society for Information Science 48, 7. p. 644-655.

Salminen, A., Lyytikäinen, V. & Tiitinen, P. (2000a), Putting documents into their work context in document analysis, Information Processing and Management 36, 4. p. 623-641.

Salminen, A. (2000b). *Methodology for document analysis.* In Kent A. (Ed.) *Encyclopedia of Library and Information Science*, vol. 67, suppl. 30. New York, Marcel Dekker Inc. p. 299-320.

Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R. (2000), *Content-Based Image Retrieval at the End of the Early Years*, IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 12. 32 p.

Veglis, A. A. (2008a), Comparison *of Alternative Channels in Cross Media Publishing,* Publishing Research Quarterly 24, 2. p. 111-123.

Veglis, A. A. (2008b), *Modeling Cross Media Publishing,* The Third International Conference on Internet and Web Applications and Services, 8.-13.6.2008 Athens, Greece, IEEE Computer Society Press, p. 267-272.

## APPENDIX 1: CrossMedia - Executive summary

The Cross Media Publishing project (CrossMedia) will create an innovation framework for cross media publishing, which develops solutions for both content and media production with interfaces to user generated content (editorial systems, supply chain management) and for consumption of cross media services within the field of learning. The project addresses all the major functions of evolving mass media: news and information provision, education and entertainment. It is also well in line with the TULEVA Strategy Outlook for Media Industry specifying the most important R&D fields until 2014.

The overall business oriented objective is to develop infrastructures which support interoperability in cross-media publishing and through these promote new business opportunities and new forms of interactions on local and communal levels.

As research goal, the interoperable infrastructures will be defined by achieving the following:

- extensible metadata systems within next generation editorial systems, which unite professional and user needs
- business models and demonstrator system for a flexible supply chain allowing capacity and resource management in production and distribution of customized media products and services
- cross media solutions for life-long learning that enable the tailoring of the personal learning environment, while increasing the use of hybrid media

The viewpoint in this project is that of the professional media – how to adapt to the changing environment, customer driven markets and increasing customization and personalization, geo-awareness and interactivity of the cross media services.

## APPENDIX 2: Metadata archive sample specification

---
1) Sample types
---
We request samples from metadata attached to articles and images. The samples need not contain the actual published content (i.e. article text or image file), only metadata fields and metadata values.

Articles
The article metadata samples should contain metadata that is explicitly connected to text at an article level. For example, metadata attached to archived pdf-files of published articles which include images is beyond the scope here.

Images
The samples should contain metadata that is explicitly connected to individual images.

---
2) Fields and values
---
The samples should show all inserted values and corresponding field names. All available fields should be shown in each sample, even though they might not have a value assigned to them.

The sample may be delivered in any convenient form, e.g. XML or text form.

---
3) Amount
---
We request 200 samples in total per paper/organization, 100 of each type described in 1) which means 100 instances of metadata attached to articles (text only), and 100 instances containing metadata attached to images. Additionally, if your company produces video, a sample of metadata attached to 100 video files is requested.

---
4) Time span
---
Random sample from year 2008

---
5) Deadline
---
Please send to chip.gylfe@tkk.fi by March 27th

In case of questions or problems, contact Chip Gylfe by:
e-mail: chip.gylfe@tkk.fi
tel. 044 - 335 1027

# APPENDIX 3: Questionnaire (original)

**Organisaatiosi osallistuu tutkimusprojektiin, jossa pyritään mm. parantamaan mediasisältöjen metatietoja, asiasanoitusta ja luokitteluja, joita yleensä kerätään saatekaavioihin toimitusjärjestelmässä.**

**Ole hyvä ja vastaa kyselyyn. Vastaamalla kyselyyn autat kehittämään saatekaavioita.**

Yksikkösi: _____

Työtehtäväsi: _____

**Oletko allaolevien mainintojen kanssa samaa mieltä? Ympyröi sopiva vaihtoehto (1 = täysin eri mieltä, 5 = täysin samaa mieltä). Vastaa En osaa sanoa (EOS), jos kysymys ei koske sinua.**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Eri dokumenttien saatetiedot** | | | | | | | |
| 1 | Lisään/muokkaan päivittäin itse laatimieni dokumenttien saatetietoja | 1 | 2 | 3 | 4 | 5 | EOS |
| 2 | Lisään/muokkaan päivittäin jonkun muun laatimien dokumenttien saatetietoja | 1 | 2 | 3 | 4 | 5 | EOS |
| 3 | Lisään/muokkaan päivittäin kuvien saatetietoja | 1 | 2 | 3 | 4 | 5 | EOS |
| 4 | Lisään/muokkaan päivittäin artikkelien/juttujen saatetietoja | 1 | 2 | 3 | 4 | 5 | EOS |
| 5 | Lisään/muokkaan päivittäin muuntyyppisten dokumenttien saatetietoja | 1 | 2 | 3 | 4 | 5 | EOS |
| **Saatetietojen täyttäminen** | | | | | | | |
| 6 | Saatetietojen täyttämiseen on annettu riittävän tarkat ohjeet | 1 | 2 | 3 | 4 | 5 | EOS |
| 7 | Minulle on selvää, miksi saatetietoja täytetään | 1 | 2 | 3 | 4 | 5 | EOS |
| 8 | Minulle on selvää, kenen tulisi täyttää tietyt saatetiedot | 1 | 2 | 3 | 4 | 5 | EOS |
| 9 | Minulle on selvää, mitä saatetietoja kuhunkin dokumenttiin tulisi täyttää | 1 | 2 | 3 | 4 | 5 | EOS |
| 10 | Minulle on selvää, milloin saatetietojen tulisi olla täytettynä | 1 | 2 | 3 | 4 | 5 | EOS |
| 11 | Saatetietokenttien merkitys on selkeä | 1 | 2 | 3 | 4 | 5 | EOS |
| 12 | Listoilta valittavien termien tai arvojen merkitys on selkeä | 1 | 2 | 3 | 4 | 5 | EOS |
| 13 | Minulle on selvää, mitä vapaatekstikenttiin tulisi syöttää | 1 | 2 | 3 | 4 | 5 | EOS |
| **Saatetietojen hyödyllisyys** | | | | | | | |
| 14 | Valtaosa kuvien saatetiedoista on hyödyllisiä | 1 | 2 | 3 | 4 | 5 | EOS |
| 15 | Valtaosa artikkelien/juttujen saatetiedoista on hyödyllisiä | 1 | 2 | 3 | 4 | 5 | EOS |
| 16 | Valtaosa muun tyyppisten dokumenttien saatetiedoista on hyödyllisiä | 1 | 2 | 3 | 4 | 5 | EOS |
| 17 | Täytetyistä saatetiedoista on hyötyä omassa työssäni | 1 | 2 | 3 | 4 | 5 | EOS |
| 18 | Täytetyistä saatetiedoista on hyötyä organisaatiossani | 1 | 2 | 3 | 4 | 5 | EOS |
| 19 | Täytetyistä saatetiedoista on hyötyä lukijoille | 1 | 2 | 3 | 4 | 5 | EOS |
| 20 | Koen, että itse täyttämistäni saatetiedoista on hyötyä | 1 | 2 | 3 | 4 | 5 | EOS |
| 21 | Tiedän, mihin täyttämiäni saatetietoja käytetään | 1 | 2 | 3 | 4 | 5 | EOS |
| 22 | Saan palautetta täyttämistäni saatetiedoista | 1 | 2 | 3 | 4 | 5 | EOS |
| **Saatetietojen ongelmat** | | | | | | | |
| 23 | Sama saatetieto täytyy syöttää useaan kertaan | 1 | 2 | 3 | 4 | 5 | EOS |
| 24 | Sama saatetieto täytyy syöttää useaan järjestelmään | 1 | 2 | 3 | 4 | 5 | EOS |
| 25 | Eri julkaisukanavia/asiakkaita varten täytetään eri saatetietoja | 1 | 2 | 3 | 4 | 5 | EOS |
| 26 | Saatetietojen täyttäminen on riittävästi automatisoitu | 1 | 2 | 3 | 4 | 5 | EOS |
| 27 | Saatetietojen täyttäminen vie liikaa aikaa suhteessa muihin tehtäviin | 1 | 2 | 3 | 4 | 5 | EOS |
| 28 | Täytettyjä saatetietoja täytyy tarkistaa/korjata myöhemmin | 1 | 2 | 3 | 4 | 5 | EOS |
| 29 | Saatetiedot ovat usein puutteelliset tai virheelliset | 1 | 2 | 3 | 4 | 5 | EOS |
| 30 | Kiireessä täytän vähemmän saatetietoja kuin tavallisesti | 1 | 2 | 3 | 4 | 5 | EOS |
| 31 | Kiireessä täytän vähemmän saatetietoja kuin pitäisi | 1 | 2 | 3 | 4 | 5 | EOS |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 32 | Täydennän täyttämiäni saatetietoja usein myöhemmin | 1 | 2 | 3 | 4 | 5 | EOS |

Kiitos vastauksistasi!

# APPENDIX 4: Questionnaire (translated)

**Your organization participates in a research project, which partial goal is to improve the metadata, keywords, and content classification that is attached to media content that commonly is collected within the editorial system.**

**Please answer this questionnaire. By doing so you aid in improving current metadata forms.**

Unit: _____

Work description: _____

**Do you agree with the statements below? Highlight the most appropriate answer (1 = totally disagree, 5 = totally agree). Answer Cannot Tell (CT), if the question does not concern you.**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Metadata in different document types** | | | | | | | |
| 1 | I add/modify metadata daily in documents that I have created. | 1 | 2 | 3 | 4 | 5 | CT |
| 2 | I add/modify metadata daily in documents that I someone else has created. | 1 | 2 | 3 | 4 | 5 | CT |
| 3 | I add/modify metadata daily in image files. | 1 | 2 | 3 | 4 | 5 | CT |
| 4 | I add/modify metadata daily in articles/text files. | 1 | 2 | 3 | 4 | 5 | CT |
| 5 | I add/modify metadata daily in other documents. | 1 | 2 | 3 | 4 | 5 | CT |
| **Metadata creation** | | | | | | | |
| 6 | I have been given sufficient instructions regarding metadata creation | 1 | 2 | 3 | 4 | 5 | CT |
| 7 | The reason for inserting metadata is clear to me | 1 | 2 | 3 | 4 | 5 | CT |
| 8 | It is clear to me who is responsible for filling in specific metadata fields | 1 | 2 | 3 | 4 | 5 | CT |
| 9 | It is clear to me, which specific metadata that needs to be filled in, and into which documents | 1 | 2 | 3 | 4 | 5 | CT |
| 10 | It is clear to me when the metadata needs to be filled in | 1 | 2 | 3 | 4 | 5 | CT |
| 11 | The meaning of the metadata fields are clear to me | 1 | 2 | 3 | 4 | 5 | CT |
| 12 | The meaning of the values available for selection in lists are clear to me | 1 | 2 | 3 | 4 | 5 | CT |
| 13 | It is clear to me what is allowed to insert into free-text fields | 1 | 2 | 3 | 4 | 5 | CT |
| **Benefits of metadata** | | | | | | | |
| 14 | The majority of image metadata is useful | 1 | 2 | 3 | 4 | 5 | CT |
| 15 | The majority of article/text metadata is useful | 1 | 2 | 3 | 4 | 5 | CT |
| 16 | The majority of other document's metadata is useful | 1 | 2 | 3 | 4 | 5 | CT |
| 17 | I personally benefit from inserted metadata | 1 | 2 | 3 | 4 | 5 | CT |
| 18 | My organization benefits from the inserted metadata | 1 | 2 | 3 | 4 | 5 | CT |
| 19 | The readers benefit from the inserted metadata | 1 | 2 | 3 | 4 | 5 | CT |
| 20 | I feel that the metadata that I create is beneficial | 1 | 2 | 3 | 4 | 5 | CT |
| 21 | I know how the metadata that I create is used | 1 | 2 | 3 | 4 | 5 | CT |
| 22 | I get feedback on the metadata that I create | 1 | 2 | 3 | 4 | 5 | CT |
| **Metadata difficulties** | | | | | | | |
| 23 | The same metadata is entered multiple times | 1 | 2 | 3 | 4 | 5 | CT |
| 24 | The same metadata needs to be entered into multiple systems | 1 | 2 | 3 | 4 | 5 | CT |
| 25 | Different metadata needs to be created for different customers or publishing channels | 1 | 2 | 3 | 4 | 5 | CT |
| 26 | The creation of metadata is sufficiently automated | 1 | 2 | 3 | 4 | 5 | CT |
| 27 | The creation of metadata is too time consuming compared to other work | 1 | 2 | 3 | 4 | 5 | CT |
| 28 | Created metadata needs to be verified/modified later on | 1 | 2 | 3 | 4 | 5 | CT |
| 29 | Metadata is often false or insufficient | 1 | 2 | 3 | 4 | 5 | CT |
| 30 | When in a hurry, I insert less metadata than usual | 1 | 2 | 3 | 4 | 5 | CT |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 31 | When in a hurry, I insert less metadata than I should | 1 | 2 | 3 | 4 | 5 | CT |
| 32 | I enhance the metadata that I create later on | 1 | 2 | 3 | 4 | 5 | CT |

Thank you for your reply!

## APPENDIX 5: Total number of rejected, missing, and “cannot tell” answers in questionnaire data

| | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | Q10 | Q11 | Q12 | Q13 | Q14 | Q15 | Q16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Rejected:** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **Missing:** | 3 | | | | 2 | | | | 1 | | 1 | 3 | | 1 | | 1 |
| **EOS:** | 6 | 7 | 4 | 4 | 15 | 2 | | | 1 | | | 9 | 10 | 19 | 14 | 41 |
| **Missing + EOS %** | 11 % | 9 % | 5 % | 5 % | 21 % | 3 % | | | 3 % | | 1 % | 15 % | 13 % | 25 % | 18 % | 53 % |
| **Total # answers:** | **71** | **73** | **76** | **76** | **63** | **78** | **80** | **80** | **78** | **80** | **79** | **68** | **70** | **60** | **66** | **38** |

| | Q17 | Q18 | Q19 | Q20 | Q21 | Q22 | Q23 | Q24 | Q25 | Q26 | Q27 | Q28 | Q29 | Q30 | Q31 | Q32 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Rejected:** | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **Missing:** | | | 1 | | | 1 | | 1 | | | 1 | 2 | 1 | | | |
| **EOS:** | 4 | 9 | 19 | 8 | 1 | 6 | 12 | 16 | 30 | 10 | 3 | 7 | 11 | 5 | 8 | 3 |
| **Missing + EOS %** | 5 % | 11 % | 25 % | 10 % | 1 % | 9 % | 15 % | 21 % | 38 % | 13 % | 5 % | 11 % | 15 % | 6 % | 10 % | 4 % |
| **Total # answers:** | **76** | **70** | **60** | **72** | **79** | **73** | **68** | **63** | **50** | **70** | **76** | **71** | **68** | **75** | **72** | **77** |

## APPENDIX 6: Translations of quotes from interviews and observations

*“jos kaikki ihmiset olisi sellaisia arkistonhoitajia että ne täyttäisi järkevästi ne laput mutta kun me ollaan kaikki omia persoonia niin se on sellaista hakuammuntaa aina sitten välillä sieltä löytää sitten mikä logiikka tässä on takana”*

*”if all people were like archivists and logically filled in those tickets but since we’re all individuals it’s a bit like shooting blanks and every now and then you find some logic behind all this.”*

*“Jotenkin en ainakaan kaipaa yhtään lisää niitä kenttiä. Kun meitä on niin erilaisia ihmisiä, joku käyttää ja toinen ei ja sitten systeemi taas hetken päästä vaihtuu. On oikeastaan ihan hyväkin että osa asioista menee henkilöltä toiselle kyselemällä.”*

*”Somehow I really don’t feel like I need any more of those fields. Because we are so different, some use them and some don’t and then the system changes. It’s actually good that some of the issues are dealt with by asking other people.”*

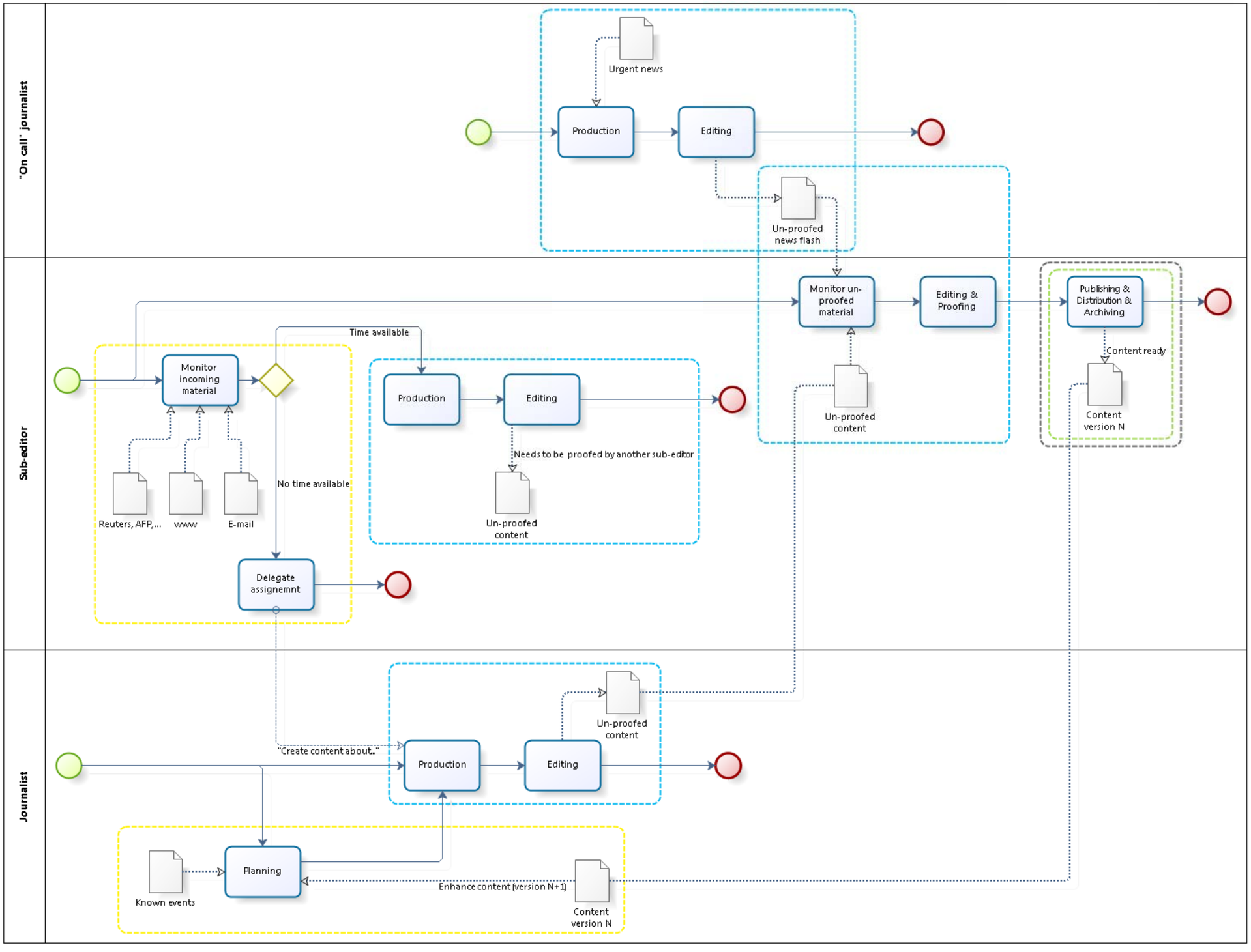
*”Mulle on epäselvää ne joita meidän ei tarvitse käyttää, eli en pysty myöskään tarkkaan perustelemaan miksi me ei niitä käytetä koska sitä ei ole.. meille on perusteltu ja selvitetty se mihin kenttiin meidän täytyy laittaa tieto, ja tiedän miksi mutta muista en tiedä.”.*

*”Those that we don’t need to use are unclear to me, and I cannot exactly specify why we don’t use them because it hasn’t.. we’ve been told and instructed about which fields we need to insert information into, and I know why but for the other fields I could not say.”*

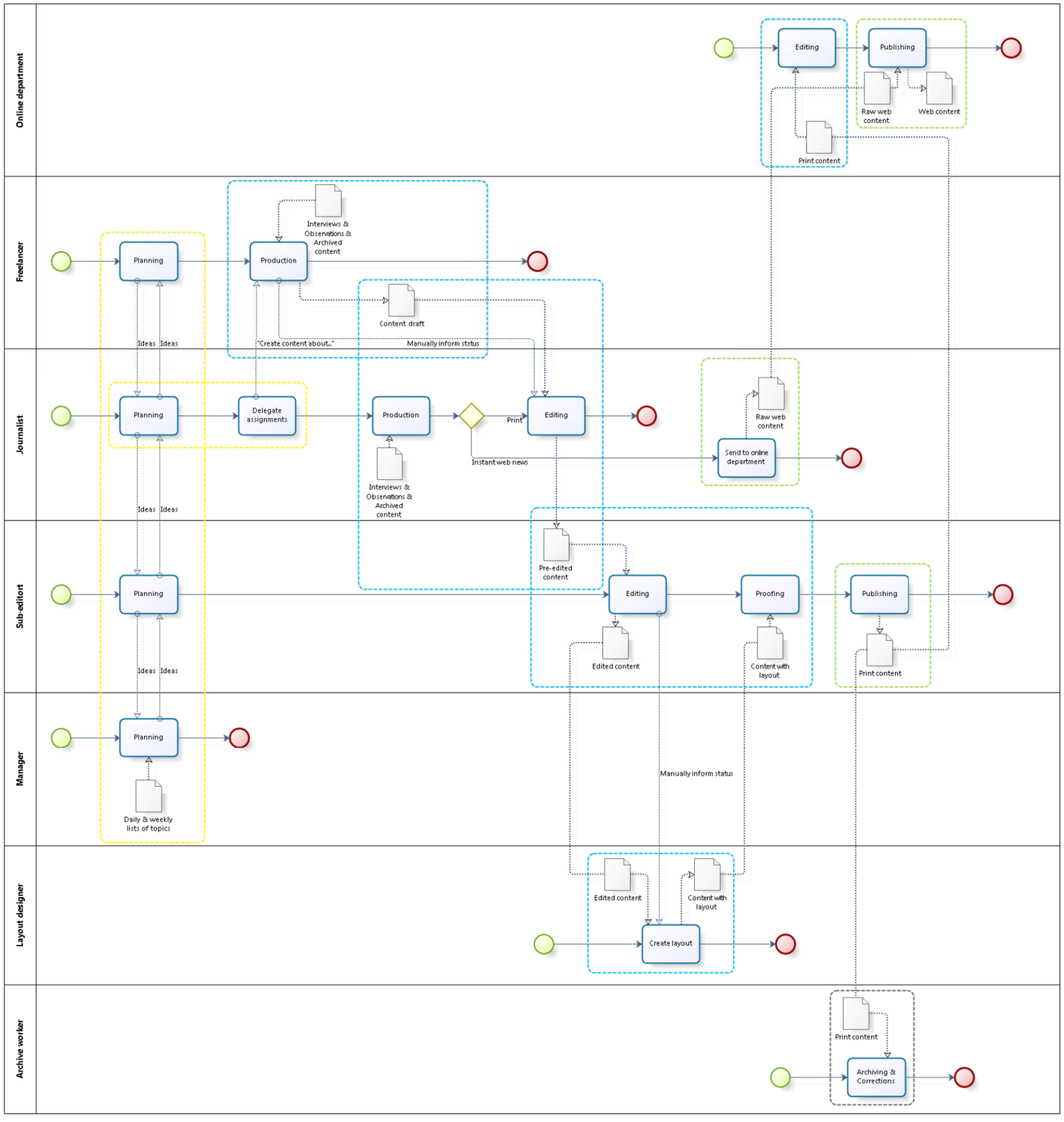
*”siellä on aika paljon mitä ei käytetä ollenkaan enää...”*

*”there is quite a lot which is not used at all anymore...”*

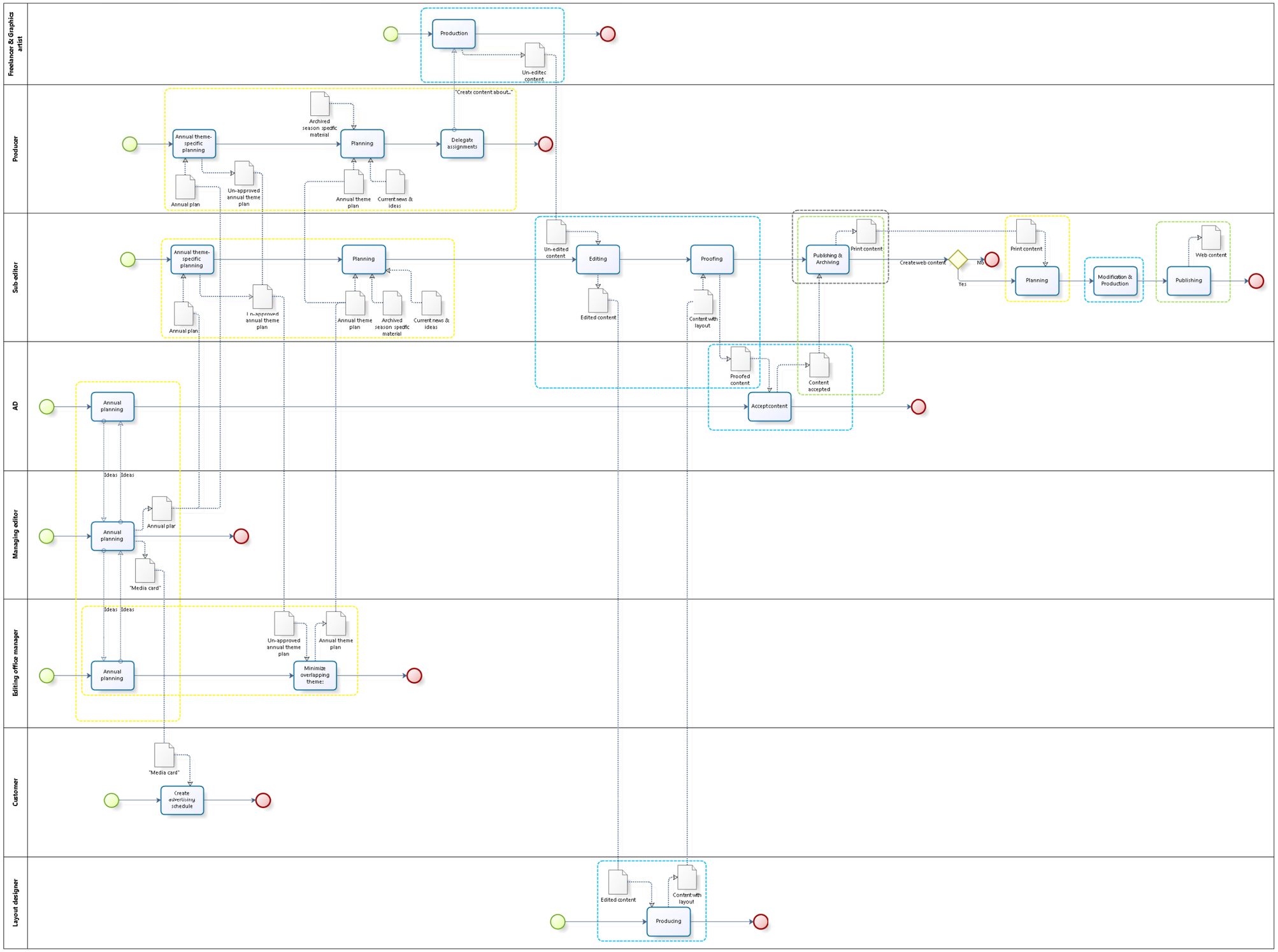
## APPENDIX 7: BPMN model of STT



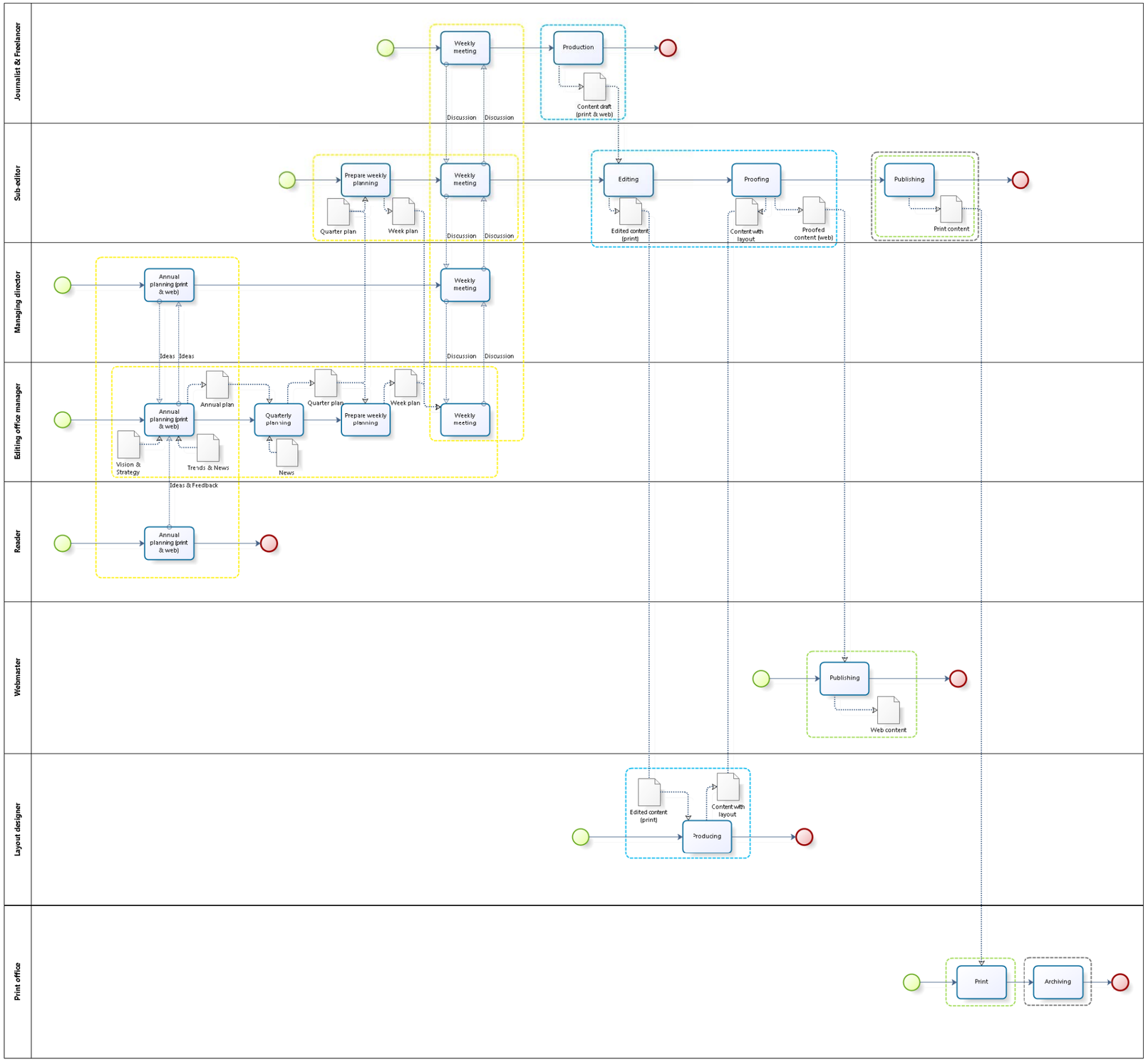
## APPENDIX 8: BPMN model of HS



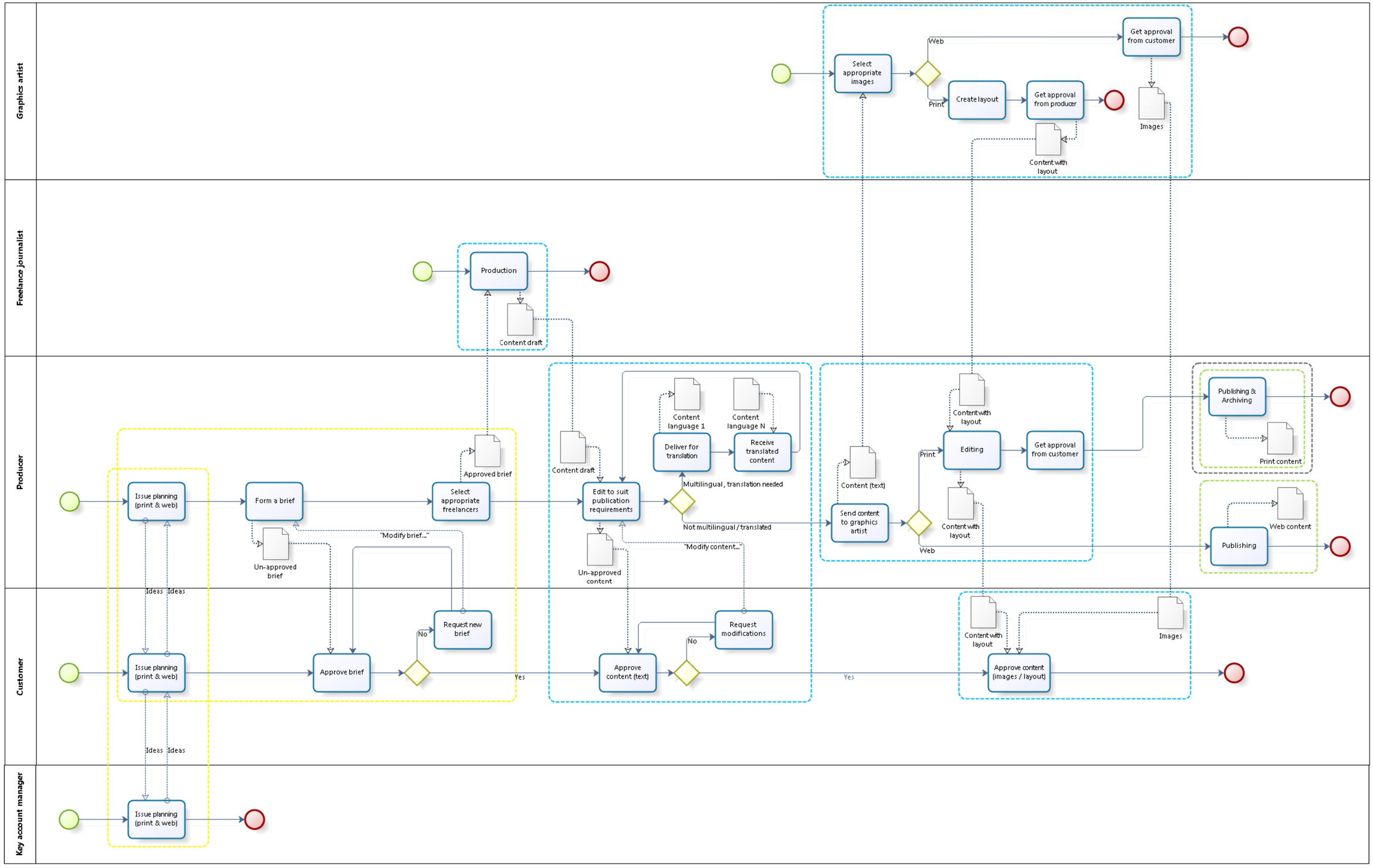
## APPENDIX 9: BPMN model of Sanoma Magazines – ET



## APPENDIX 10: BPMN model of Sanoma Magazines – MB



**APPENDIX 11: BPMN model of Sanoma Magazines – YJ**



## APPENDIX 12: Metadata categorization by Jokela (2001)

| Type | Definition | Examples |
|---|---|---|
| **Structural** | Describes the structural characteristics, the format, of the content essence, but does not contain information about what the content essence actually means. | Decoding information related to e.g. video, audio, or graphics formats. Other relevant areas include compression data, composition and synchronization information, as well as information on sequencing the content essence. |
| **Control** | Related to controlling the flow of content. It is used to determine whether the content is ready to progress in the content value chain or not. Commonly more temporary in nature, when compared to other metadata elements. | Machine control, QoS, error management, etc. |
| **Descriptive** | | |
| **Contextual** | Describes the environment and conditions of content essence and its creation. | Geospatial information, timing information, and information on the equipment used to produce the content essence. |
| **Content-based semantic** | Describes semantic qualities of the content essence answering the question what the content essence means. It is needed for the processing or usage of the content essence. It describes such qualities such as the subject, location, names, and style of the content essence. | Keywords of a news story. Also typically used in highly content-based products, such as in personalized news services, |
| | Descriptive metadata can also be used for controlling the use of content essence, and is therefore closely related to control metadata. Examples of usage areas are: intellectual property and access rights, as well as information on supported media platforms. | |

## APPENDIX 13: Metadata categorization by Gilliland-Swetland (2000)

| Type | Definition | Examples |
|---|---|---|
| **Administrative** | Metadata used in managing and administering information resources. | Acquisition information, rights and reproduction tracking, documentation of legal access requirements, location information, version control and differentiation between similar information objects, and audit trails. |
| **Descriptive** | Metadata used to describe or identify information resources. | Cataloging records, finding aids, specialized indexes, and hyperlinked relationships between resources. |
| **Preservation** | Metadata related to the preservation management of information resources. | Documentation of physical condition of resources, as well as actions taken to preserve both physical and digital versions of resources. |
| **Technical** | Metadata related to how a system functions or metadata behave. | Hardware and software documentation, digitization information, tracking of system response times, and authentication and security data. |
| **Use** | Metadata related to the level and type of use of information resources. | Exhibit records, use and user tracking, content re-use and multi-versioning information. |

## APPENDIX 14: Metadata categorization by Mauthe & Thomas (2004)

| Type | Definition | Examples |
|---|---|---|
| **Content-related** | Describes the actual content or subject matter. | Description of video and image content. |
| **Material-related** | Describes available formats, encoding parameters, and recoding specific information. | Video or image format, their compression type and amount etc. |
| **Location-related** | Describes location and amount of copies, condition of carriers, etc. | Could specify if there is a "master" version as well as copies made of it. |

## APPENDIX 15: Metadata categorization by Boll et al. (1998)

| Type | Definition | Examples |
|---|---|---|
| **Media type-specific** | Different media types have specific needs concerning metadata. | Texture of images, frequencies in audio, etc. |
| **Media processing-specific** | Describes functions for processing specific media. | Search and retrieval functions, functions for directing transfers, performance measuring to insure sufficient QoS, etc. |
| **Content-specific** | Metadata that is derived solely by content and independent of media type. | Text, images and video may all share the same content specific metadata. The content is simply presented with different media types. |

## APPENDIX 16: Properties and definitions from IPTC Photo Metadata Core Schema 1.1 (Anon 2008c)

| | |
|---|---|
| City (legacy) | Name of the city the content is focussing on -- either the place shown in visual media or referenced by text or audio media. This element is at the third level of a top-down geographical hierarchy. |
| Country (legacy) | Full name of the country the content is focussing on -- either the country shown in visual media or referenced in text or audio media. This element is at the top/first level of a top-down geographical hierarchy. The full name should be expressed as a verbal name and not as a code, a code should go to the element "CountryCode" |
| Country Code (legacy) | Code of the country the content is focussing on -- either the country shown in visual media or referenced in text or audio media. This element is at the top/first level of a top-down geographical hierarchy. The code should be taken from ISO 3166 two or three letter code. The full name of a country should go to the "Country" element. |
| Description | A textual description, including captions, of the item's content, particularly used where the object is not text. |
| Headline | A brief synopsis of the caption. Headline is not the same as Title. |
| Intellectual Genre | Describes the nature, intellectual, artistic or journalistic characteristic of a item, not specifically its content. |
| Keywords | Keywords to express the subject of the content. Keywords may be free text and don't have to be taken from a controlled vocabulary. Codes from the controlled vocabulary IPTC Subject NewsCodes must go to the "Subject Code" field. |
| Province or State (legacy) | Name of the subregion of a country -- either called province or state or anything else -- the content is focussing on -- either the subregion shown in visual media or referenced by text or audio media. This element is at the second level of a top-down geographical hierarchy. |
| Scene Code | Describes the scene of a news content. Specifies one or more terms from the IPTC "Scene-NewsCodes". Each Scene is represented as a string of 6 digits in an unordered list. |
| Subject Code | Specifies one or more Subjects from the IPTC Subject-NewsCodes taxonomy to categorise the content. Each Subject is represented as a string of 8 digits in an unordered list. |
| Sublocation (legacy) | Name of a sublocation the content is focussing on -- either the location shown in visual media or referenced by text or audio media. This location name could either be the name of a sublocation to a city or the name of a well known location or (natural) monument outside a city. In the sense of a sublocation to a city this element is at the fourth level of a top-down geographical hierarchy. |
| Date Created | Designates the date and optionally the time the intellectual content was created rather than the date of the creation of the physical representation. If no time is given the value should default to 00:00:00. |
| Description Writer | Identifier or the name of the person involved in writing, editing or correcting the description of the content. |
| Instructions | Any of a number of instructions from the provider or creator to the receiver of the item. |
| Job Id | Number or identifier for the purpose of improved workflow handling. This is a user created identifier related to the job for which the item is supplied. |
| Title | A shorthand reference for the item. Title provides a short human readable name which can be a text and/or numeric reference. It is not the same as Headline. |
| Copyright Notice | Contains any necessary copyright notice for claiming the intellectual property for this item and should identify the current owner of the copyright for the item. Other entities like the creator of the item may be added in the corresponding field. Notes on usage rights should be provided in "Rights usage terms". |
| Creator | Contains the name of the person who created the content of this item, a photographer for photos, a graphic artist for graphics, or a writer for textual news, but in cases where the photographer should not be identified the name of a company or organisation may be appropriate. |
| Creator's Contact Info | The creator's contact information provides all necessary information to get in contact with the creator of this item and comprises a set of sub-properties for proper addressing. |
| Creator's jobtitle | Contains the job title of the person who created the content of this item. As this is sort of a qualifier the Creator element has to be filled in as mandatory prerequisite for using Creator's Jobtitle. |
| Credit Line | The credit to person(s) and/or organisation(s) required by the supplier of the item to be used when published. This is a free-text field. |
| Rights Usage Terms | The licensing parameters of the item expressed in free-text. |
| Source | Identifies the original owner of the copyright for the intellectual content of the item. This could be an agency, a member of an agency or an individual. Source could be different from Creator and from the entities in the CopyrightNotice. |
| Contact Information Details structure | A generic structure providing a basic set of information to get in contact with a person or organisation. |
| Address {contact info detail} | The contact information address part. Comprises an optional company name and all required information to locate the building or postbox to which mail should be sent. To that end, the address is a multiline field. |

| | |
|---|---|
| City {contact info detail} | City {contact info detail} |
| Country {contact info detail} | The contact information country part. |
| Email address(es) {contact info detail} | The contact information email address part. |
| Phone number(s) {contact info detail} | The contact information phone number part. |
| Postal Code {contact info detail} | The contact information part denoting the local postal code. |
| State/Provice {contact info detail} | The contact information part denoting regional information like state or province. |
| Web URL(s) {contact info detail} | The contact information web address part. Multiple addresses can be given, separated by a comma. |