Libraries, archives and museums collect and manage paintings, statues, books, songs, films and other objects. To manage their collections, cataloguers describe the objects’ attributes such as title, maker, subjects etc. The descriptions serve two purposes: 1) to create collection access: to retrieve objects, group objects and find similar objects and 2) to signify objects: to contextualize them, relate them to other objects or to concepts (e.g. communistic art), to find structure in multiple objects (type, trend, style, group of makers).

The Netherlands Institute for Sound & Vision is the Dutch national audiovisual archive. Digitization reduced the cost of production and storage of digital material, resulting in a huge increase in stored and inflowing material (approximately 40,000 hours of new material was stored in 2014). Next to the increase in inflowing material, digitization is a game changer for Sound & Vision. The Dutch public broadcasters transformed to a completely digital infrastructure in 2006. In this landscape, catalogue users, such as professional broadcasters, expect (annotated) new material to turn up in the online catalogue on the day of the broadcast.

Manual annotation is a time consuming activity. At Sound & Vision 42 cataloguers are employed. They manage to catalogue roughly one third of all yearly inflowing material, with the final description being ready roughly two months after the broadcast. The rest has only the metadata with which it was delivered to the archive, which most of the time is its title plus technical broadcast information such as date, time and channel of broadcast.

Sound & Vision cataloguers use two indexing tools: iMMiX and GTAA. iMMiX is Sound & Vision’s archiving system. iMMiX is split into two linked parts: the iMMiX metadata system stores the metadata and "Het Digitale Archief" (the digital archive) stores the essence (the primary audiovisual material). The iMMiX metadata contains a.o. technical information (e.g. length, broadcast date), content descriptors (e.g. keywords, locations, persons, makers, producers) and digital rights information. Most of the content descriptors, such as keywords, locations, persons, maker and producer, are restricted to S&Vs in house thesaurus named GTAA (GTAA is a Dutch acronym for “Gemeenschappelijke Thesaurus Audiovisuele Archieven”\textsuperscript{13}). The GTAA contains about 160,000 terms, organized in 6 disjoint facets: Subjects, Genres, People, Names\textsuperscript{14}, Makers and Locations.

\textsuperscript{13}Stands for Common Thesaurus [for] Audiovisual Archives.
\textsuperscript{14}For Organization, Groups and Other Names.
Terms from all facets of the GTAA may have Related Terms and Scope Notes, but only Keywords and Genres can also have Use/Use for and Broader Term/Narrower Term relations, the latter organizing them into a set of hierarchies. Terms from the four other facets are alphabetical flat lists, sometimes associated with ScopeNotes (textual comments about the use of the given term).

This makes our research question: **How can cultural heritage cataloguers be supported automatically during the creation and enrichment of catalogue annotations?**

Our hypothesis is that the structured background information, stored in thesauri such as the GTAA and in metadata (e.g. in iMMiX), is useful for supporting cataloguers during the process of indexing. We take an engineering approach and design (automatic) processes and build prototypes. These are subsequently evaluated during experiments.

In Chapter 2: *Web-Based Thesaurus Browsing* we described the development and evaluation of a thesaurus browser. In the thesaurus browser we tried to visualize the different aspects of the GTAA thesaurus: facets, the terms, their relations, their hierarchies and the facet categories. We tested and adapted our design during two iterations of users performing an indexing task with the thesaurus browser. Sound & Vision’s cataloguers were used to quite basic tools for finding index terms and were dazzled by the complexity of the browser interface. They were used to alphabetical search and therefore we gained most performance by optimizing the alphabetical search part of the browser. Navigating the hierarchical representation was apparently of much use for finding terms. For disambiguation however, showing the terms in the hierarchy is a quick mean to determine if the term is the right concept. We noticed a difference of strategies between the subjects from Sound & Vision (experts of the thesaurus content) and from broadcasting corporations. The broadcasters were more eager to start the search for a term by browsing. Thus a browsing facility could be helpful to the daily users of Sound & Vision’s archives.

The cataloguers were in general positive about the use of the designed tool in their daily work. This was not only apparent from the questionnaire, but also from the fact that, based on the results of this study, Sound & Vision incorporated aspects of the thesaurus browser in their archiving process.

Chapter 3: *Automatically Deriving Annotations* described a prototype which we designed to use structured background information in combination with analysis of contextual information of archival objects for the automatic generation of annotation suggestions. The architecture for the automatic creation of annotations in an audiovisual archive is sketched in Figure A.1.

The prototype analyzed contextual information, in this case being four web site texts which belonged to two audiovisual broadcasts, and used the information in the GTAA thesaurus to rank the terms extracted from the contextual texts. The prototype’s outcome were six types of ranked annotations. This outcome was evaluated by exact matching against a) one set of expert descriptions and b) one set of descriptions made by nine cataloguers in a user study. The exact matching against the expert description was the classic (information retrieval) evaluation. The evaluation against
the annotations of a group of nine cataloguers gave insight into the inter cataloguers agreement. Both these evaluations showed that sometimes annotations can be right but not an exact match (for example when The Caucasus was suggested when both Chechnya and Dagestan appeared in the human annotations\(^\text{15}\)). To address this, we introduced semantic evaluation in Chapter 4.

In Chapter 4: Evaluation of Automatic Annotations we fully implemented the automatic annotation prototype from Chapter 3 and focussed on the evaluation of its outcomes\(^\text{16}\). We created a corpus of 258 broadcasts for which we derived automatic annotations. On these we performed a classic information extraction evaluation. This evaluation however, when analyzed into more detail, showed limitations. These limitations were explored in two ways: first by changing the evaluation to semantic evaluation, second by introducing the idea of serendipitous browsing.

The looser semantic evaluation judges terms within one thesaurus relation to catalogue keywords as correct as well. It is able to capture small differences which lead to unwanted wrong evaluations. For example: the set automatic annotations ministers, pigs pest, agriculture, ministries and pigs, is completely wrong when compared exactly against the manually assigned keywords pest, cattle, vaccinations and veterinary diseases. When evaluated semantically however pigs pest, agriculture and pigs are judged

\(^{15}\)The Caucasus is the mountain range between Chechnya and Dagestan where the program makers filmed their program.

\(^{16}\)We implemented three different ranking algorithms for step III in Figure A.1.
correct. Changing to a semantic evaluation is a good strategy when evaluating automatic indexing tools, given the cataloguers indexing task and the conceptual modeling of the GTAA thesaurus.

With serendipitous browsing (automatic and manual) annotations are used for finding interesting related documents. It showed that the manual annotations and the automatic annotations have the same value for finding interesting related documents.

In Chapter 5: Automatic Keyword Suggestion we separately studied different realizations and different inputs of four components of the automatic annotation process.

In Section 5.2 we studied the ranking algorithms. We saw that the ranking algorithms influence the outcome of the keyword suggestions. Our best algorithm was the TF.RR algorithm which uses the thesaurus relations and frequency information. It equaled the classic TF.IDF algorithm without the need for a background corpus when evaluated classically. It added to TF.IDF by creating homogeneity in its results and in the semantic evaluation.

From the different types of structured background information, which we studied in Section 5.3, the handmade GTAA was clearly the best. It outperformed the co-occurrence network derived from the iMMiX catalogue descriptions.

In Section 5.4, we gathered content information on the indexable objects either with ASR or via context documents. It showed that both the ASR and the context documents are valuable sources and complementary sources of information, but that the ASR might not yet be as good as context documents for the derivation of annotations.

Even though we dedicated Chapter 4 to three ways of evaluation of automatically generated keyword suggestions, we studied the interaction between automatic annotation and query expansion in Section 5.5\textsuperscript{17}. It showed that automatic annotations can be complementary to manual annotations in a retrieval setting.

In Chapter 6: Documentalist Support System we integrated automatic annotations and the visualization of the information environment in the documentalist support system (DocSS). The DocSS allows a cataloguer to inspect all the information relevant during the cataloguing of an individual object. It shows information on the catalogue collection level, in related documents and their associated texts, it shows information on the individual objects level, showing the context documents, generates annotations in these documents (spotted GTAA terms) and creates ranked sets of annotations extracted form these documents.

Although this prototype was not implemented in the working archive, it offered interesting possibilities for semantic browsing and searching the rich semantic graph of Sound & Vision’s catalogue. The automatic linking to thesaurus terms and similar documents, the creation of navigational links showing the information environment with information scent [136] and its interactive nature, gives many ideas for new highly interactive user interfaces for information interaction.

In the final research chapter, Chapter 7: User Evaluation of Manual and Automatic Annotations, we studied the differences and overlap between manual and automatic annotations.\textsuperscript{17} Yet another evaluation of automatically generated annotations was performed in Chapter 7, where user feedback was given on lists of automatically generated annotations in a user experiment.
annotation in a user study. In this evaluation we had different user groups judge the perceived value of both manual and automatic annotations. The three groups were cataloguers, catalogue users and interested layman-users.

When we looked at the overlap between the manual and automatic annotations in Section 7.4, we saw that 40% of the good annotations was generated both by the cataloguers and the automatic techniques. The manual and automatic annotations differ in the remaining 60%. 40% of the good annotations was only generated by the automatic techniques. The last 20% was only generated by the human cataloguers. Even though the automatic annotations contained the most good annotations, they got accompanied by a very large set of bad annotations, whereas the manual annotations were almost all good.

This does not mean however, that the automatic results are useless for a cataloguer. Its results where the worst for the makers annotation, but even with only a third of the makers suggestions correct, offering these to cataloguers changes their task from picking 7 makers from 18,000 to picking 6 of these from a set of 20. This may increase the number of annotations and the working speed of cataloguers. A study into the effect of automatic pre-descriptions on the working speed of a human annotator shows a change in working speed of +25% to +35% [46].