3.1. Publishable summary

Live video content is increasingly consumed over IP networks in addition to traditional broadcasting. The move to IP provides a huge opportunity to discover what people are watching in much greater breadth and depth than currently possible through interviews or set-top box based data gathering by rating organizations, because it allows direct analysis of consumer behavior via the logs they produce. The ViSTA-TV project proposes to gather consumers' anonymized viewing behavior and the actual video streams from broadcasters/IPTV-transmitters to combine them with enhanced electronic program guide information as the input for a holistic live-stream data mining analysis: the basis for an SME-driven market-place for TV viewing-behavior information.
First, ViSTA-TV employs the gathered information via a stream-analytics process to generate a high-quality linked open dataset (LOD) describing live TV programming. Second, combining the LOD with the behavioral information gathered, ViSTA-TV is now in the position to provide highly accurate market research information about viewing behavior that can be used for a variety of analyses of high interest to all participants in the TV-industry. This generates a novel, SME-driven market place for TV viewing-behavior data and analyses. Third, to gather anonymized behavioral information about viewers not using our IPTV-streams ViSTA-TV will focus its second year to employ the gathered information to build a recommendation service that exploits both usage information and personalized feature extraction in conjunction with existing meta-information to provide real-time viewing recommendations.
Commercially, the revenues gathered in the market research activity will cross-subsidize the production of the open-sourced LOD stream.
These results are made possible by scientific progress in data-stream mining consisting of advances in (1) data mining for tagging, recommendations, and behavioral analyses and (2) temporal/probabilistic RDF-triple stream processing.

In the past video content was consumed through broadcast TV delivered through the air or by cables. Broadcasters had information – both meta-data and the actual video stream – about their own programming as well as coarse usage information supplied by survey companies such as Nielsen or GfK.[1] Cable providers viewed themselves as pure bundling and transportation channels whilst viewers used secondary sources to inform themselves about the desired programming sources. Fueled by Internet TV (IPTV) offers of telecommunication giants, cable providers, or pure online providers such as YouTube or Hulu as well as digital rental companies such as NetFlix and iTunes, people increasingly watch video content delivered over IP networks. The technological issues arising from IPTV as well as the opportunities for interactive, personalized TV have been amply investigated in EU-projects such as NoTube. However, the use of behavioral information accruing during the process of viewing live IPTV in combination with the actual video streams and the electronic program guide (EPG) information as a foundation for a market research data marketplace has been largely overlooked.

[1] http://nielsen.com, http://gfk.com

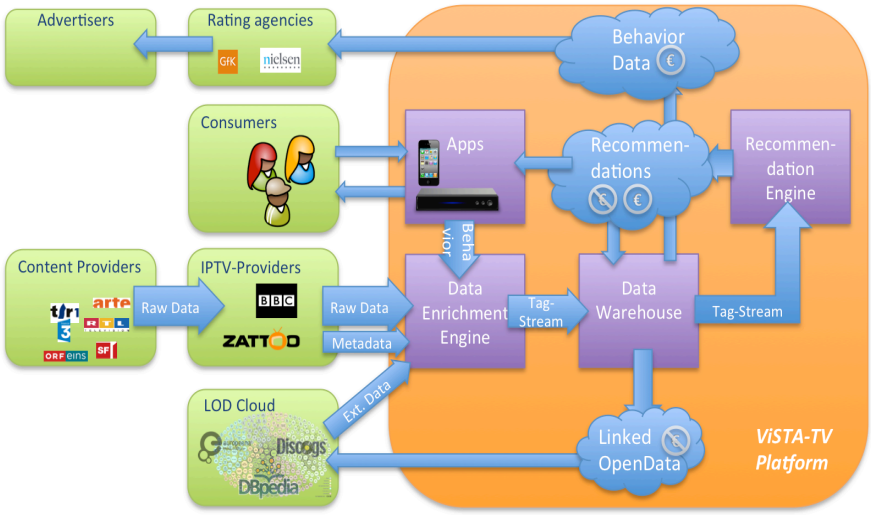

Figure 1: ViSTA-TV Overview (logos are owned by respective companies)

The ViSTA-TV project considers IPTV as a two-way channel, where the viewer can take advantage of the video streams whilst the ViSTA-TV platform employs behavioral information about the viewer gathered by IPTV transmitters to improve the experience for all participants in the TV supply chain (see Figure 1). Specifically, ViSTA-TV gathers consumers' viewing behavior and the actual video streams from broadcasters and IPTV providers to combine them with enhanced EPG information as the input for a holistic live-stream analysis. The analysis – comprising of personalized feature construction, supplementary tag generation, and real-time recommendation generation – provides the input for a triplication procedure that generates Data in the Resource Description Framework (RDF) – a data format for storing typed graphs [2]– and enables the reuse of the resulting data for three different data pools:

(1) Linked Open Data as a Basis for Analytic Processing,
(2) Viewing Behavior Data, and
(3) Recommendation Data.

Each of these data pools with its production pipeline provides the foundation for a data marketplace with its own commercial as well as societal rationale such as bootstrapping innovative applications relying on TV data, selling services relying on detailed viewing behavior analysis beyond today's capabilities, or helping viewers finding the shows that best match their interest at any given point in time.
The heart of the ViSTA-TV platform is an online analysis system that operates on streams of content data in real time. This system is developed on the basis of the industry-proven streaming platforms STORM.[3] We extended existing methods for feature extraction from content streams, as well as methods for querying and matching events on streams of background information to make them stream-ready. Last but not least we developed

[2] RDF models typed graphs as a collection of edges, each of which is described a 3-tuple (called a triple) consisting of the source node identifier, the edge type, and the target node identifier or a literal.
[3] http://storm-project.net

methods for real-time recommendation based on complex event processing and hybrid user/show profiles. For learning and evaluating the corresponding data models we rely upon existing offline machine learning methods, in particular agglomerative and two-way hierarchical clustering.