Challenges in Heterogeneous Web Data Analytics - Case Finnish Growth Companies in Social Media

Jaakko Salonen
Tampere University of Technology
Korkeakoulunkatu 3
33720 Tampere, Finland
jaakko.salonen@tut.fi

Jukka Huhtamäki
Tampere University of Technology
Korkeakoulunkatu 3
33720 Tampere, Finland
jukka.huhtamaki@tut.fi

Ossi Nykänen
Tampere University of Technology
Korkeakoulunkatu 3
33720 Tampere, Finland
ossi.nykanen@tut.fi

ABSTRACT
Diverse data about various phenomena are implicitly available in the modern web. In particular websites categorized as social media provide rich and heterogeneous data about various entities such as people, corporations, brands as well as their properties and relationships. An analyst who seeks to leverage this diverse data is faced with the challenge of integrating and making sense of a set of heterogeneous data sources. In this paper, we provide an introduction and a problem statement for heterogeneous web data analytics. To further highlight and discuss practical challenges, we introduce a case study of Finnish growth companies in social media. Instead of a purely data-driven approach, the presented approach is rooted in the idea that an analyst can actively participate in the data collection and integration process, while the process can still retain repeatability and transparency. The key contribution of this paper is the statement of the challenges related to heterogeneous web data analytics.

Categories and Subject Descriptors
H.1.2. [**Information systems**]: Models and principles – *User/Machine Systems*.

H.2.5. [**Information systems**]: Database management – *Heterogeneous Databases.*

General Terms
Algorithms, Design, Experimentation, Human Factors, Languages

Keywords
Social media, Data analysis, Social network analysis, Crawling, Linked data, Big data

1. INTRODUCTION
Diverse data on various phenomena are implicitly available in the modern web, providing potentially a valuable source of information for an analyst. In particular, services categorized as social media provide rich data sets about various entities, such as people, corporations and brands, as well as their properties and relationships. Similar information can be found from individual websites as well, yet not in such abundance.

Academic MindTrek 2013, October 1-4, 2013, Tampere, FINLAND.

Information about these entities can be readily retrieved from individual sites and services both in structured and unstructured formats. Facebook and Twitter, for example, provide well-documented APIs for accessing data in structured formats. Structured information is often available on arbitrary websites as well, either 1) directly, as embedded in content using formats like Microformats, RDFa[1], and Open Graph Protocol or 2) indirectly in referenced resources in formats including RSS, Atom, and JSON[2].

Some pieces of information are only available in unstructured formats including plain text, non-semantic HTML or even URL fragments. Various computational methods, such as text analysis and data mining, may need to be applied in an attempt of making sense of this data (e.g. [1]).

Even though not often part of a theoretical data analysis frameworks, a real, complementary issue in web data analysis is data access. APIs often provide clean contracts for data access, including information on pricing and access limits, making the data access a manageable issue in these cases. However, extraction of ad-hoc web content has proven to be a much more complex issue. For instance, an analyst may choose to use his or her own credentials on automating data retrieval directly from authorized sites, such as Facebook. Even if formally conforming to the terms of use, such access methods may be perceived by the service provider as malicious, resulting in access restrictions or even bans. We see a clear need for web data access methodology that allows one to extract data in a legal fashion, both on principle as well as perceivably.

In the long run, the issues of data integration may be perhaps resolved by codifying, structuring, serving and extracting data according to the specifications and practices as endorsed by the Semantic Web and Linked Data communities ([2], [3], [4], [5]). However, we see that intermediary practices and tools are needed for today's analysts to work with, and to fill the gap between current web of data and a visionary Web of fully structured data.

We have particularly identified the need of further work on practices and tools for dealing with heterogeneous datasets. With *heterogeneous dataset* we refer to a dataset that has one or more of the following properties:

1. **Multi-sourced**. The dataset has been collected from multiple data sources.
2. **Multi-structured.** The dataset mixes together with multiple levels of structuring (unstructured, semi-structured or structured data)

[1] Resource Description – in – attributes

[2] JavaScript Object Notation

3. **Multi-schematic.** The dataset contains two or more pieces of data that conform to different schemata.

To address these issues, this paper presents a model for collecting and combining heterogeneous web data into integrated data models, in a fashion that allows us to work with various levels of data (structured, semi-structured, unstructured) with potentially restricted data access. The data collection, refinement and integration model is based on an iterative process that is actively driven by an analyst. Instead of a purely data-driven approach, the process is rooted in the idea that, in the spirit of interactive computing, the analyst can actively participate in the data collection and refinement process, while the process can still retain repeatability and transparency.

The rest of the paper is structured as follows: In section 2, background on the topic and related work are presented. Section 3 proceeds to present a preliminary process model we utilize in heterogeneous data analytics. In section 4, an analysis case regarding Finnish growth companies in social media is described. In section 5 results of the analysis case are presented. In the last section we discuss future work and conclude.

2. BACKGROUND AND RELATED WORK

In this section we briefly outline background for our study and related work in terms of related research disciplines. Firstly, we describe the background and motivation for our study within the context of innovation ecosystems research. We then proceed on presenting closely related work on the fields of social media analytics, scientific data analysis and visual analytics.

2.1 Motivation and Challenges

Motivation for heterogeneous web data analytics in this particular study lies in our current research efforts of studying innovation ecosystems in Reino research project.

Structured, curated data is generally considered as a primary source of data for studying the innovation ecosystems. Yet, open access to online data has made a wealth of data widely available and permits researchers to leverage it for insights about the emergence and evolution of innovation ecosystems.

The use of online data sources has been successfully demonstrated for descriptive innovation ecosystems mapping [6]. Socially constructed data has been used in a similar fashion to create network maps for mobile ecosystem [7], EIT ICT Labs inter-city mobility [8], as well as the national innovation ecosystem [9]. Social media data is seen as a potentially valuable source of complementing data for analysis [10].

Along with the changing nature of innovation activities as well as the availability of online data or even big data, researchers developing more representative indicators for innovation seek to apply secondary data, i.e. data collected from online sources and social media [10]. Whereas individual sets of secondary data have already been successfully used to provide new insights on business and innovation ecosystems, linking the datasets has proven to be a difficult challenge [7].

Innovation ecosystems research is an example of a research domain in which harnessing heterogeneous mixes of web data sources is attempted. Curated data is available in structured or semi-structured formats. On the other hand, some further insight maybe gained by consolidating it with data available in the web in unstructured formats. The challenge can be posed as follows: how can we effectively retrieve data and combine it in a way that provides us with meaningful representations of innovation ecosystems?

2.2 Related Social Media Analytics Research

Individual blogs and collections of blogs including the blogosphere (i.e. the aggregate of all public blogs) have been readily studied. Current directions and opportunities for social media analytics on the blogosphere have been outlined, with topics including 1) seeking relevance of blogosphere discussion topics, 2) understanding and discovering influence and authority, 3) detecting sentiments, and, finally, 4) discovering emerging topics [11]. Additionally, texts in individual blogs as well as collections of blogs have been analyzed for blog profiling, text classification, comment spam detection, blog sentiments, comments, search behavior and opinion retrieval [12].

From individual social media sites, Twitter in particular, has been widely studied. Macro-level properties of Twitter in its early years have been reported in quantitative studies [13][14]. Individual Twitter messages (tweets) and their linguistic and semantic properties have been investigated, often with emphasis on sentiment analysis including analysis of mood and emotions [15][16][17], political sentiments [18], and consumer opinions on brands and products [19][20]. In addition, a case study regarding semantics and context of Twitter discussion has been reported [21].

Indicating the importance of this topic, a number of commercial products for social media analytics are available, including HootSuite[3], Infergy[4] and SAS® Social Media Analytics[5].

2.3 Related Scientific Data Analysis and Visualization Research

Our particular motivation on innovation ecosystems can be viewed as a process of scientific data analysis, in which interactive visualization plays a key role. We also recognize the various architectures and models for interactive visualizations.

A characterization and requirements for scientific data analysis process have been reported by Springmeyer, Blattner and Max [22]. The paper reports a decomposition of the scientific data analysis process along with its five functional requirements. The requirements are: 1) facilitation of active exploration, 2) capturing the context of analysis, 3) linking materials from different stages of a study, 4) minimizing unnecessary or distracting navigation requirements and 5) providing computer support for culling large data sets. [Ibid.]

An often cited information visualization reference model has been reported by Card, Mackinlay and Schneiderman [23]. The reference model recognizes visualizations as adjustable mappings from data to visual form. The mappings are: 1) data transformations from raw data to data tables, visual mappings of data tables to visual structures, and 3) mapping of visual structures to various views via view transformations. [Ibid.]

Complementary to the reference model, data-flow architecture has been reported by Abram, and Treinish [24]. Further, Jankun-

[3] https://hootsuite.com/

[4] http://infegy.com/

[5] http://www.sas.com/software/customer-intelligence/social-media-analytics.html

Kelly, Ma, and Gertz [25] characterize a model and a framework for visualization exploration.

Heer and Shneiderman [26] present a taxonomy of tools that support the fluent and flexible use of visualizations. The taxonomy consists of 12 task types grouped into three high-level categories: (1) data and view specification (visualize, filter, sort, and derive); (2) view manipulation (select, navigate, coordinate, and organize); and (3) analysis process and provenance (record, annotate, share, and guide). [Ibid.]

A closely related practical implementation for networked data analysis is the Orion system [27]. In Orion, an analyst is provided with a linker tool that allows flexible construction of network models from multi-table schema based input data. The system allows analysts to also specify filters and data aggregation operations [Ibid.]. We see Orion as a particularly good fit to be used as a later-stage tool in a heterogeneous data analysis process, given that we can provide it with sufficiently well consolidated input data.

3. PROCESS MODEL

Based on our past experiences on web data analysis as well as drawing implicitly from the related analytics research, a process model for data analysis has been formulated (see Fig. 1).

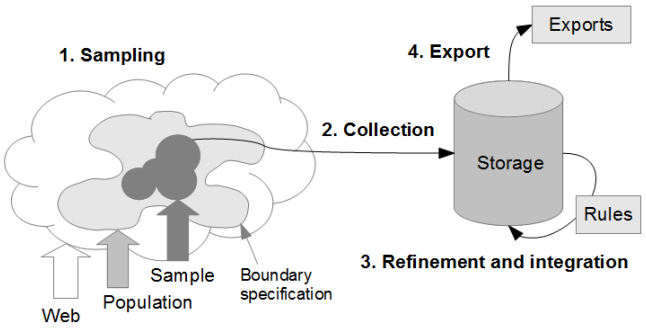

Figure 1: A preliminary model for collecting and processing heterogeneous web data.

The process can be briefly described as an iterative, four step process in which an analyst may interact at any point. The four steps in the process are: data 1) sampling 2) collection, 3) refinement and integration and 4) export.

In the *sampling step*, the initial points of interest in source data are specified from a given set of interesting web content (i.e. *population*). The desired source data points are specified as a *sample*. At its simplest, a sample may simply consist of a list of interesting URL addresses that an analyst sees as relevant. More sophisticated sampling methods may involve crawling web data further from the initial addresses. In such cases the actual sampled data is the result of the crawling process that has been performed according to given *boundary specifications* (for example: "crawl all related resources that are maximum of two steps away from a list of seed URLs").

The notion of sampling is important in several respects. Firstly, online data often only provides us with external representations of various phenomena. For instance, if one seeks to study growth companies, only a limited amount of information about them has been made available online. Secondly, access to information in available online resources may be restricted, in particular due to data access issues or due to poor quality of unstructured data. And, finally, it is sometimes sufficient to only obtain a sample of all available data for specific analytical purposes.

In the *collection step,* the sampled data is retrieved by whatever automated means necessary. In practice, data collection may be implemented with web crawlers, scrapers or other data access tools that retrieve content based on specifications specified as the sample. Data collection tools may implement intermediary data storages or caches. These mechanisms include, but are not limited to: 1) general purpose (HTTP) request caching, 2) resource level (URL/URI-specific) caching, and 3) document and data object persistence (flat files, databases).

The important point is that the collection process needs to be data-driven: if an analyst needs to repeat the data collection and caching, that should be possible simply by re-running the collection process, optionally by changing the desired sample. The running of the collection process results in the (re)generation of a dataset.

In the *refinement and integration step,* the collected datasets are further processed to allow creation of one, inter-connected result dataset. Some data transformations may already occur as part of the collection step. The focus is in this step is on transforming the collected dataset into a coherent data model, as required by the analysis task in hand.

As data is often collected from multiple sources (APIs, sites, services), the resulting, aggregated dataset has usually been modelled according to multiple schemata. By introducing a top-level data model in this step, these schemata could be combined. In addition, various transformation or normalization rules can be applied.

In some cases, the collected data may be simply broken. Issues with character sets, for example, will often end up causing problems. For such cases, data refinement and integration step may be used to provide custom data patches or fixes. In all these data modification cases, it is important that the process is transparent and repeatable. For this reason, we note that the enrichment and refinement step should provide analysts with explicitly notion of what specific transformations (*Rules*) have taken place.

Finally, in the *export step*, a subset of the collected, enriched and refined dataset is exported for further data processing. The export step outputs one or more representations of the data from dataset (*Exports*). The exports may represent data either directly or in stored formats. Optionally the export step may apply visual mapping rules, which effectively transform the data into visual structures (cf. [23]).

An important notion in the presented process is the ability for an analyst to both interactively manage each step in the process as well as the ability to iteratively repeat the process as a whole. For instance, an analyst may choose to initially select a specific sample, and later adjust the sample according to experiences from latter steps with the initial sampling. Similarly, any other part of the process may be adjusted while interacting with the data, just keeping in mind that it is important to keep the data transformation process as a whole as transparent and repeatable as possible.

4. CASE: FINNISH GROWTH COMPANIES IN SOCIAL MEDIA

In order to demonstrate the practical problems related to data analysis in the scope of the presented context, a data analysis case investigating the social media presence of some Finnish growth companies was chosen.

4.1 Case Study Description

In the case study, we investigate the social media presence of the companies that were active participants of Tekes Young Innovative Companies (IYC) program during June 2013. Tekes[6] is a Finnish funding agency for technology and innovation. Tekes YIC program specifically provides support to young innovative companies for fast international growth [28].

A list of companies that are currently funded is publicly available on Tekes website [29]. We chose this full list of currently participating companies as the case study dataset for our study.

Data from the website listing was collected and curated into a data table representing a list of social media resources related to the given company. The curated data table consisted of the following columns: company name, company (primary) product name, Twitter username, Facebook username, Homepage URL, Blog URL and Blog Feed URL.

We further proceeded to implement a data collection and analysis process for this dataset as described in the next subsection.

4.2 Implementation

4.2.1 Data Collection and Sampling

The data for each item (company) was sampled from three channels (Twitter, Facebook and Blog feeds) according to the model presented in Figure 2.

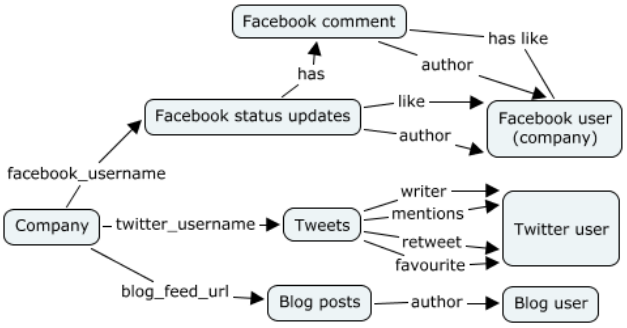

Figure 2: Data collection model.

The data collection was performed as a combination of three separate data processing scripts: Twitter importer, Facebook importer and Blog feed importer. Each individual script was required to collect information in several consecutive data processing steps from several web locations, which resulted in a network of information.

The *Blog feed importer* was implemented to iterate through all the last blog posts, as available in the blog feed provided by the company website. In addition to blog post data, some feeds provided information about blog post author (username).

The *Facebook importer* was implemented by accessing data via Facebook Graph API[7] . Facebook usernames were used as an entry point to the API. The graph API was further navigated to import all latest status updates for each user, as well as related comments, likes and users assigned to them.

The *Twitter importer* was implemented by accessing Twitter's API with Tweepy[8]. For each company all latest tweets available via the API were retrieved along with information about users related to the tweets as direct mentions in the tweet. In order to limit amount of data collected and number of requests required to send to Twitter API, retweets and favorites were not included.

From a technical viewpoint, the importers were implemented with Python and modeled in pandas[9] as DataFrame objects. Several Python and data processing utility libraries were used in the process, including requests[10], and feedparser[11]. DataFrame objects produced by the importers were stored in multiple formats for interactive analysis and further processing (HTML, pandas DataFrame save method).

4.2.2 Data Refinement and Integration

After data collection, a data refinement processing step was implemented in two steps: 1) network model generation and 2) network merging.

In *network model generation*, each individual data import was processed and converted into a two-mode network. Nodes in the generated network represented either individual actors (companies or individual persons) or individual resources (blog posts, Facebook wall posts, tweets).

In *network merging phase*, the individual networks were merged into a multi-data source based composite network. Individual nodes in the network were merged simply based on their labels (Twitter user title, Facebook username, Blog username).

A set of programmatic as well as manually curated data normalization rules was created for the merging. Entity names were programmatically normalized by stripping out various commonly used company name suffixes ("Oy", "Ltd", "AB", etc.). Additionally, a manually curated set of normalization rules was applied. These rules were interactively developed during the data analysis process. For instance if there would be slightly different fashions in how a company name was written, a normalization rule could be written to merge these names into a single entity.

4.2.3 Data Exports and Visualizations

The data was exported in Graph Exchange XML Format (GEXF). Both intermediary, single data-source based networks (Facebook, Twitter, blogs) as well as a refined composite network were exported. The exported data were then analysed and visualized with Gephi [30].

5. RESULTS

Total of 88 companies were collected from the list of companies currently funded by Tekes [29]. For each company individual

[6] http://www.tekes.fi/

[7] http://developers.facebook.com/docs/reference/api/

[8] http://tweepy.github.io/

[9] http://pandas.pydata.org/

[10] http://python-requests.org

[11] http://pythonhosted.org/feedparser/

Twitter and Facebook account names as well as blog URLs were manually curated to form an initial, sampling dataset. After curating the implemented data importers were ran to gather the data. Data from each importer was sampled by choosing only maximum of 10 of the latest items available (Blog post, Facebook status update, Twitter updates).

Overall statistics of the gathered data are the following:

- 250 individual blog posts by 67 unique authors
- 401 Facebook wall posts by 42 unique authors and 1277 unique users with likes or follow-up comments
- 494 Tweets by 52 unique authors and 244 mentions of distinct users

Additionally, a composite network aggregating together all three gathered data sets was created. The composite network contained:

- 1669 nodes from Facebook (wall posts, users)
- 765 nodes from Twitter (tweets, users)
- 317 nodes from blogs (blog posts, users)
- 37 nodes (users) that co-existed two sources
- 1 node that co-existed in all three data sources

A visualization of the composite network was created with Gephi and is illustrated in Figure 3. Nodes from separate data sources are illustrated with colour encoding (light blue for Facebook, purple for Twitter and green for blogs, and other colours for nodes with multiple data sources).

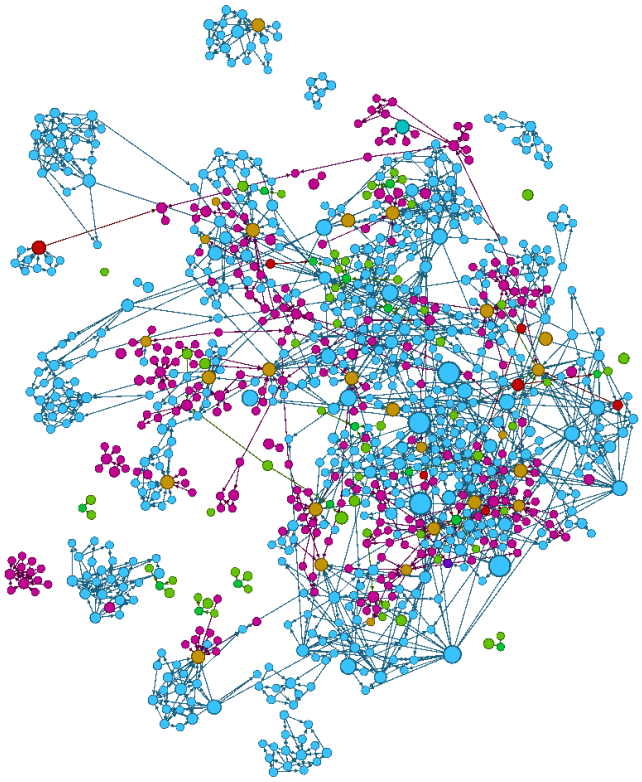

Figure 3: Visualization of the composite network.

The distribution of edge types was following (Fig. 4): 51.53% likes (Facebook), 30.25% author of, 9.53% user mentions (Twitter), 6.5% commented by (Facebook) and contributed to blog (2.18%).

5.1 Discussion

The case data analysis process presented here raised multiple challenges, both regarding the analysis of web data in general as well as the integration of heterogeneous data in general. We name these challenges as six issue topics. The topics are: 1) data access, 2) data structuring, 3) data validity and precision, 4) data integration and modeling, 5) data bias and sampling, and finally 6) legal and contractual issues. We will briefly discuss each of these topics separately.

5.1.1 Data Access Issues

Practical issues regarding data access were raised during the analysis several times. A data source may either become temporarily unavailable or access to it may be intentionally restricted. Twitter API, for example, has strict rate limits. In simple cases, these rate limits can be easily dealt with by restricting the number of requests a data access script makes. However, if one wishes to sample large amounts of data from Twitter, it may become significantly slow or even impossible. For instance we attempted to collect precise data about retweets and favorites, but quickly found out that analysis loop became too slow with these data collections enabled. Additional, Twitter API provides only a limited access to favorites and retweets: only first 100 retweets can be accessed and favorites need to be parsed on per-user, not on per-tweet, basis [31].

A preliminary workaround for data access issues is, in general, to use more intermediary cache and storage mechanisms that optimize API quota usage. Alternatively, various third party data sources with less strict access limits could be used.

5.1.2 Data Structuring Issues

Poorly structured data was an issue especially regarding blog data analysis. For instance some web sites had news or blog-like content, but did not provide an alternative structured representation (RSS or Atom feeds). In some cases the data could have been accessed from inline semantic markup formats (RDFa, microformats), which we didn't implement, but most of the time the data did not encode any easily interpretable semantics. Lack of data structuring was also the reason we completely excluded blog comments from our data collection process.

One possible way to deal with the data structuring issues is to research and implement data scrapers that utilize soft computing methods for detecting and scraping specific kind of content.

5.1.3 Data Validity and Precision Issues

Even well-structured data can be problematic if it is not valid or sufficiently precise. In terms of data validity, we generally have assumed that collected metadata is valid. For instance, if Facebook or blog claims a specific author, we generally have assumed it to be true. A malicious data provider could easily distort our analysis by providing invalid author data. As such, we clearly see data validity as a key concern in producing valid insights.

Another related problem is the one with data precision. For instance in blogs, the *author* metadata field, may only provide first name of a given author or, instead, just company name. In these cases, it is difficult or impossible to track down the exact name of the actual content author. Similarly, multiple authors may have been involved in writing of blog content, which may not be accurately reflected in metadata. As such data precision is an additional, related challenge.

At some levels, issues with data precision or validity may be addressed by data patching rules. Also, in some cases, precision may be improved by intelligently fusing together more information from the data collection context. In Twitter, for instance, a common practise is to include more precise author information in Twitter account's description ("Tweets by @username"). Yet, we see that further research needs to be done in order to better address these issues.

5.1.4 Data Integration and Modelling Issues

On a higher level, data integration can be seen as a challenge in merging together multi-schemata data. Simple example is when one wishes to consolidate data where data has been encoded with various syntaxes. For instance, a person may be encoded as "*Firstname Lastname"* in other and *"Lastname, Firstname"* in the other dataset. In more complex cases, the analyst will find out himself or herself dealing with intricacies of practical linguistics, including detection and management of homonyms, synonyms, pseudonyms, acronyms and abbreviations.

We see that in order to solve challenges regarding data integration, both sufficient data modelling and application of domain expertise are required, not to mention a suitable standard representation for the data. A data model used in integration should allow modelling of the source data in sufficiently detailed fashion. For instance, instead of only collecting names or titles of specific resources, a data collection model should optimally provide sufficient context for more detailed script and rule-based data integration. Specifically URL addresses and other identifiers as well as other contextual metadata should be included in the models.

Yet the models alone are not often sufficient. The actual data integration may require application of domain specific expertise in the formalization of various integration rules and scripts. For instance, in a specific data analysis case, an analyst may specify the rules to represent all legal entities related to a brand as a single entity (e.g. Company Oy, Company Ltd., Company Inc. is just "Company"). Yet in other analysis cases, however, distinction between the specific legal entities may be needed (three different companies).

Another aspect of the integration challenge is providing the analyst with sufficient automation. Even the most precise dataset does not provide much value, if compiling it is either too slow or too expensive.

Scalability of data integration practices may be challenged by the explosion in number of possible data sources. For instance, in our case study, only three types of data or data sources were used (Facebook, Twitter and blogs via web feeds). In practice an analyst would benefit from the ability to add new data sources. In our practical case this would mean the ability to consolidate data from other websites and social networks, notably LinkedIn, Wikipedia, YouTube, Google+ and Xing, as well as from curated datasets like Innovation Ecosystems Network datasets (IEN Growth, IEN Startup, etc.). As the number of data sources increases, also increases the challenge of dealing with overlapping and ambiguous information.

One way to overcome issues in data integration is to use socially constructed datasets. For business ecosystems research, Crunchbase[12] provides one such data source. Another potential for solving issues in data integration is the application of semantic modelling and computing for deriving required expert knowledge from formalized data representations.

5.1.5 Data Bias and Sampling Issues

For more exploratory data analysis tasks, it may be acceptable to simply work in a fashion where data is sampled in an ad-hoc fashion. However, for some specific data analysis tasks, it may be important to determine whether the sampled data is quantitatively sufficiently representation from the population of the available data, or any bias potentially present in the data is within an acceptable range. If sample does not represent the whole population, the validity of the conclusions based on the data is compromised.

We suspect that much of the data extracted from the web in an ad-hoc fashion is for practical reasons quite biased. Firstly, it is simply easier for an analyst to extract and incorporate well-structured data into a dataset, and discard poorly structured data. In addition, consolidated data may not be well balanced between different data sources and thus, provide biased view to the phenomenon as a whole.

Data bias can be practically seen in the context of our case study. For instance, Facebook provides rich information about wall posts (likes, comments, etc.) that is easily accessible via Facebook Graph API. In Twitter, on the other hand, rich information about tweets (retweets, favourites) is difficult to access due to API usage restrictions. As a result, the composite network model in our case study has a richer set of relationships for Facebook data than for Twitter. Without better knowledge, an analyst may conclude that, for example, Facebook has more importance in social media presence, even though there might be a bias due to data access issues.

One final relevant aspect in managing the scope and bias of data is boundary specification. If we analyse networks of data, one way to limit data sampling is to a set a maximum distance (e.g. maximum of *n* steps away from seed nodes). If we mix together a heterogeneous set of networked data, this brings us the challenge of measuring if the distances in the networks are in fact comparable. For instance some networks may represent only the tightest bounds as edges, while other networks may use edges to represent any kinds of associations. Thus, sampling individual networks based on boundary specifications before merging them together may yield different results than first merging the datasets and then sampling it as a whole.

One potential way to manage data bias and work around the sampling issues would be the adoption of big data analytics tools and processes (see e.g. [32], [33]). Instead of giving strict rules and boundary specifications for sampling, one could leverage cheap and cost-effective data warehousing tools and collect data in a greedy (or "magnetic") fashion [32]. This, however, would in turn introduce biases of its own, and possibly blur the boundaries of the dataset.

5.1.6 Legal and Contractual Issues

From legal and contractual point of view, application of multi-source data is often non-trivial. In our analysis case, we are integrating data from multiple sources, including Facebook, Twitter and dozens of individual blogs and websites. Facebook and Twitter provide explicit information on terms of service, including information and licensing on content re-use [34] [35]. For individual blogs, on the other hand, such information is rarely

[12] http://www.crunchbase.com/

available. Consequently, it is difficult to define precisely how the legal terms, if available in the first place, should be integrated to allow an analyst to modify, use and redistribute the collected datasets as well as its various exports.

One way to work around with legal issues multisource data would be to prefer using data sources and data access methods like APIs that provide clear contractual frameworks. Moreover, the data collection process should allow collection and tracking of licensing information, available either as metadata or as external documents such as terms of service. Finally, we wish to point out that integration of multi-license data is a non-trivial task for which further legal work and research are likely needed. The task is further complicated by the fact that both the national and the international legal systems face the obvious difficulties in predicting the dynamics of new technologies. This establishes a slowly evolving legal legacy setting which is quite hard to be interpreted by developers.

6. CONCLUSIONS

In this paper, we have outlined potential benefits and challenges that are related to the analysis of heterogeneous web data. Further, we have briefly introduced background to the topic as well as introduced the context of our previous work regarding especially regarding innovation ecosystems analysis.

The paper proceeded to present a preliminary data analysis process model that allows an analyst to run the analysis in a data-driven fashion while retaining transparency and repeatability of the process. Further, we present a relevant case study that focuses on analyzing social media presence of young innovative Finnish growth companies that are funded by Tekes. A custom made data analysis tools and processes were implemented for the case study and were presented in this paper. The results of the data analysis from the case study were presented and evaluated and discussed. Finally, with the help of the case study, we have highlighted some of the current challenges and opportunities of the analysis of heterogeneous web data.

We see that potential for new insights exists. In the context of innovation and business ecosystem analysis, the ability to address the given challenges would allow an analyst to better analyze and cross-reference social media and online presence for a set of given companies. As the modern-day innovation activities are more user-centric than before and concentrating e.g. on combining existing solutions into new services rather than developing new technology, having access to user-level data is seen to be highly valuable [10]. Moreover, open innovation, co-creation and ecosystemic innovation in general happens more likely between companies than in R&D departments of individual organizations. This also insist accessing new kinds of online data sources.

Several interesting lines of research exist for future work. Firstly, more data sources could be included in the data set, most notably LinkedIn and Google+ network data. Secondly, data collection methods for existing data sources could be further improved as well, including crawling of blog comments, Twitter retweets and favorites and collecting Facebook friendship networks. Thirdly, data merging models could be further improved. For one, our current model does not allow making distinction between homonymous entities. This could be partly resolved by collecting per-service unique usernames and company identifiers instead of current model of using only textual titles in data merging. Fourthly, improvements in modeling of aggregate data models could be investigated. Finally, on a higher level, a more comprehensive review of literature regarding existing data refinement and integration methods would be helpful in providing a solid basis for further development. Also, open reference datasets and more research on methods for scientifically validating the various models and sampled data sets are needed, to establish a commonly agreed frame of reference for rigorously evaluating and comparing the different heterogeneous data management approaches.

7. ACKNOWLEDGMENTS

This research is sponsored by Tekes – the Finnish Funding Agency for Technology and Innovation (Project "Reino"; Relational Capital for Innovative Growth Companies) and was done in collaboration with the Innovation Ecosystems Network (http://www.innovation-ecosystems.org/).

8. REFERENCES

[1] Norvig, P. 2009. Natural Language Corpus Data. In Segaran, T., Hammerbacher, J. (Eds.) *Beautiful Data*, O'Reilly Media, 2009, 219-242.

[2] Berners-Lee, T. 2009. *Linked Data – Design Issues*. Retrieved May 19, 2013, from http://www.w3.org/DesignIssues/LinkedData.html

[3] Sauermann, L, and Cyganiak, R. (Eds.) 2008. *Cool URIs for the Semantic Web*. W3C Interest Group Note 03 December 2008. Retrieved June 25, 2013, from http://www.w3.org/TR/cooluris/

[4] Berrueta, D., Phipps, J. (Eds.) 2008. *Best Practice Recipes for Publishing RDF Vocabularies*. W3C Working Group Note 28 August 2008. Retrieved June 25, 2013, from http://www.w3.org/TR/swbp-vocab-pub/

[5] Connolly, D., (Eds.) 2007. *Gleaning Resource Descriptions from Dialects of Languages (GRDDL)*. W3C Recommendation 11 September 2007. Retrieved August 2, 2013, from http://www.w3.org/TR/grddl/

[6] Rubens, N., Still, K., Huhtamäki, J., Russell, M. G. 2010. Leveraging social media for analysis of innovation players and their moves. Stanford University.

[7] Basole, R.C., Russel, M.G., Huhtamäki, J., Rubens, N. 2012. Understanding Mobile Ecosystem Dynamics: A Data-Driven Approach. *Proceedings of the 2012 International Conference on Mobile Business (ICMB 2012)*. Delft, Netherlands, Jun. 2012, pp. 17–28.

[8] Still, K., Huhtamäki, J., Russell, M. G., Rubens, N. 2012. Transforming Innovation Ecosystems through Network Orchestration - Case EIT ICT Labs. *Proceedings of the XXIII ISPIM Conference – Action for Innovation: Innovating from Experience*, Barcelona, Spain, June 17-20, 2012.

[9] Huhtamäki, J., Russell, M. G., Still, K., Rubens, N. 2011. A network-centric snapshot of value co-creation in Finnish innovation financing. *Open Source Business Resource*, March 2011, pp. 13-21.

[10] Still, K., Huhtamäki, J., Russell, M. G., Rubens, N. 2012. Paradigm shift in innovation indicators: From analog to digital. *Proceedings of the 5th ISPIM Innovation Symposium*, Seoul, Korea.

[11] Melville, P., Sindhwani, V., Lawrence, R. D. 2009. *Social Media Analytics: Channeling the Power of the Blogosphere for Marketing Insight*.

[12] Mishne, G. A. 2007. Applied text analytics for blogs. Dissertation. University of Amsterdam, Faculty of Science. From http://dare.uva.nl/document/46517

[13] Java, A., Song, X., Finin, T., Tseng, B. 2007. Why we twitter: understanding microblogging usage and communities. *WebKDD/SNA-KDD '07 Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*.

[14] Kwak, H., Lee, C., Park, H., Moon, S. 2010. What is Twitter, a social network or a news media? In *WWW'10 Proceedings of the 19th International Conference on World Wide Web*.

[15] Bollen, J., Mao, H., Pepe, A. 2011. Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, 17-21 July 2011, Barcelona, Spain

[16] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R. 2011. In *LSM'11 Proceedings of the Workshop on Languages in Social Media*, 30-38

[17] Go, A., Bhayani, R., Huang, L. 2009. Twitter Sentiment Classification using Distant Supervision. Retrieved from http://cs.wmich.edu/~tllake/fileshare/TwitterDistantSupervision09.pdf

[18] Tumasjan, A., Sprenger, T. O., Sandner, P. G., Welpe, I. M. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pp. 178-185.

[19] Jansen, B. J., Zhang, M., Sobel, K., Chowdury, A. 2009. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, Vol. 60, Issue 11, pp. 2169-2188, November 2009.

[20] Pak, A., Paroubek, P. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, May 19-21 2010, Valletta, Malta.

[21] Russell, M., G., Flora, J., Strohmaier, M., Pöschko, J., Perez, R., Rubens, N. 2011. Semantic Analysis of Energy-Related Conversations in Social Media: A Twitter Case Study. *International Conference on Persuasive Technology (Persuasive 2011)*.

[22] Springmeyer, R. R., Blattner, M. M., Max, N. L. 1992. A Characterization of the Scientific Data Analysis Process. In *Proceedings of the 3rd conference on Visualization '92*. IEEE Computer Society Press Los Alamitos, USA, 235-242.

[23] Card, S. K., Mackinlay, J., and Schneiderman, B. (Eds.) 1999. *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann Publishers.

[24] Abram, G., Treinish, L. 1995. An extended data-flow architecture for data analysis and visualization. In *Proceedings of the IEEE Conference on Visualization*, 263-270

[25] Jankun-Kelly, T. J., Ma, K.-L., Gertz, M. 2007. A model and framework for visualization exploration. *IEEE Transactions on Visualization and Computer Graphics* 13(2): pp. 357-369; http://dx.doi.org/10.1109/TVCG.2007.28

[26] Heer, J., Shneiderman, B. 2012. Interactive Dynamics for Visual Analysis. *ACM Queue*.

[27] Heer, J., Perer, A. 2011. Orion: A System for Modeling, Transformation and visualization of Multidimensional Heterogeneous Networks. *IEEE Visual Analytics Science & Technology (VAST)*, 2011.

[28] Tekes. 2013. Funding for young innovative companies (YIC). Online. Retrieved 31 May 2013, from http://www.tekes.fi/about/niy

[29] Tekes. 2013. Usein kysyttyä; Linkit rahoituksen piirissä 31.3.2013 olevien yritysten sivuille. Online. Retrieved 31 May 2013, from http://www.tekes.fi/info/niy/usein+kysyttya

[30] Bastian, M., Heymann, S., Jacomy, M. 2009. Gephi: an open source software for exploring and manipulating networks. *International AAAI Conference on Weblogs and Social Media*.

[31] Twitter. 2013. REST API v1.1 Resources. Online. Available at: https://dev.twitter.com/docs/api/1.1 (retrieved 19.6.2013)

[32] Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J. M., and Welton, C. 2009. MAD skills: new analysis practices for big data. *Proceedings of the VLDB Endowment*, *2*(2), 1481–1492.

[33] O'Reilly Media 2012. *Big Data Now – Current Perspectives from O'Reilly Media*, 2012 Edition.

[34] Facebook. 2012. Terms of Service. Retrieved from https://www.facebook.com/legal/terms

[35] Twitter. 2012. Terms of Service. Retrieved from https://twitter.com/tos