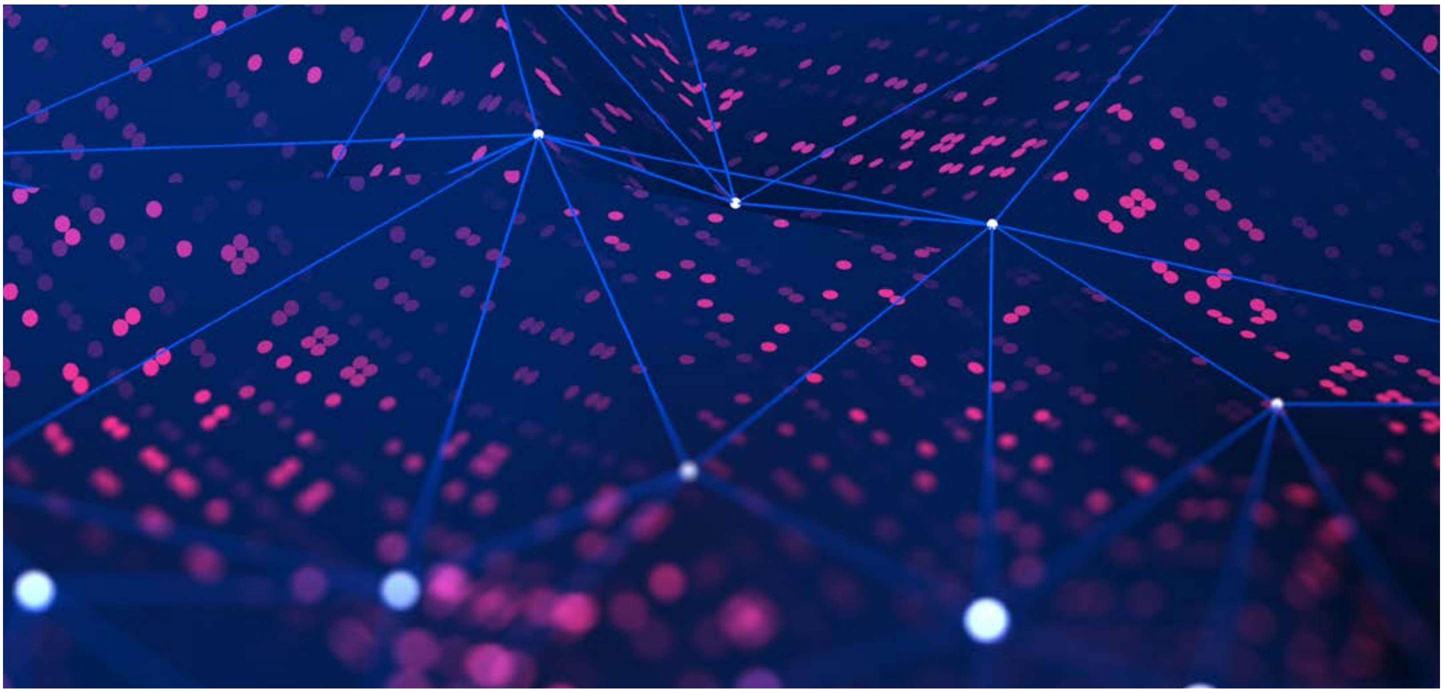
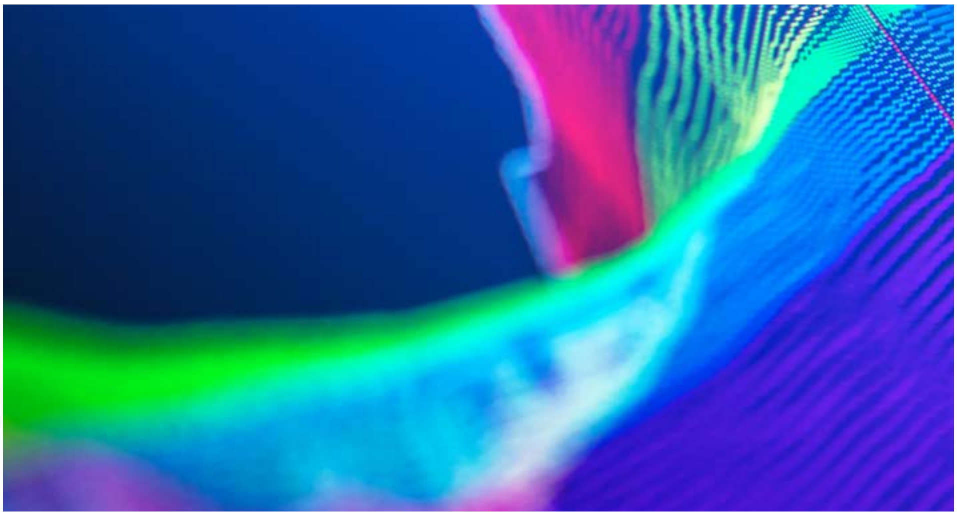
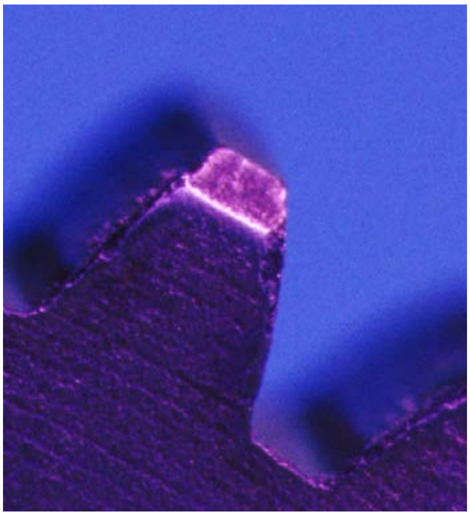
ESOMAR

Briefing questions

when considering tools and services for unstructured data – text, images, audio, and video







Briefing questions

when considering tools and services for unstructured data – text, images, audio, and video

Introduction

In 2012, ESOMAR published 24 Questions to Help Buyers of Social Media Research. These questions were intended to help users of social media research consider issues that might influence whether a social media listening tool was fit for the purpose of a particular research objective, whether qualitative, quantitative, or both. The questions were designed to help users gain a better understanding of the services being offered and ensure that what they received from a social media data provider was what was expected.

Over the intervening years a great deal has changed in terms of the types of data available for analysis, the sources for such data, the ways in which researchers acquire and analyze it, the technologies used, the industry players, and the regulatory environment, to name a few. Of special note is the interest in moving beyond text to include the broader category of unstructured data (text, images, audio, and video) and the expansion of potential sources beyond social media to include, for example, survey open ends, focus group transcripts, call center interactions, and more. At the same time, the software tools for analyzing these types of data have grown in number and capabilities. The purpose of this document is to update ESOMAR's guidance to better reflect current practice in market, opinion, and social research and data analytics.

Company profile and capabilities

1 What is the company's core business – the services offered, and verticals served?

This will help you form an opinion about the relevant experience of the provider given the specific types of unstructured data you plan to analyze, its sources, and the knowledge domain(s) referenced in the data.

2 What are the typical deliverables?

Providers typically offer a range of deliverables from raw output to client-ready graphics. You should consider carefully how you plan to use the output, the skills of your team, your client's expectations, and your ability to present your findings before deciding on the deliverables that will be most useful to you.

3 How is pricing determined?

Pricing can vary widely and be based on any number of factors including the amount of data collected and analyzed, how it is harvested, the deliverables, frequency of reporting, etc.

4 Are there case studies that can be shared?

Reviewing use cases that address the specific business decision to be made and showing how the results were used may help you to gain assurance that the solution is fit to purpose.

Data sources and types

5 What data sources does the company rely on?

There are many different sources of unstructured data and it is important that the provider you choose have access to and experience with those that are most important to your research given the geographic coverage you require, the types of data you want included, and the topic(s) being studied.

6 How does the company gather the data?

The provider may harvest the data directly from the Web, use a third-party service, or take delivery from you or your client. Or, it may use a combination of methods. Regardless of the method(s) used, your primary concerns should be how well it matches your specific data requirements and whether the provider has the legal authority to access and process the data. Regarding the latter, there may be limitations on harvesting data from sites based on a site's terms of use or local intellectual property laws.

7 Does the company provide historical data from its sources?

You may want to analyze the data from an historical perspective and so it is important to know whether historical data is available, and the specific sources and time periods covered. Some social media monitoring vendors only provide forward harvesting from the day of subscription confirmation.

Software design and capabilities

8 What types of unstructured data analysis is the software capable of producing?

Text, images, audio and video can be harvested from the web or taken from other sources. Does the software offer the capability of data harvesting or should the user find the data from other sources and upload it to the software? Can the software provide buzz (word counts), sentiment, specific emotions, and semantic (topics) analysis? How extensive is the inductive topic analysis provided – if any? Can it analyse images for objects, brand logos and theme? Can it analyse audio for sentiment, emotions and topics?

9 Does the software use machine learning or an engineered approach to produce the analyses?

An engineered approach involves the manual (or machine assisted) creation of dictionaries, lexicons, thesauri, rules, and taxonomies. A machine learning approach may be supervised, with models built from annotated training data, or unsupervised, for instance generating a topic model based on a notion of statistical similarity. Further, "Active learning is a special case of semi-supervised machine learning in which a learning algorithm is able to interactively query the user (or some other information source) to obtain the desired outputs at new data points," and reinforcement learning implies that the machine will weigh outcomes in order to learn continuously.

If an engineered approach is taken, the user should understand whether it is possible to customize the language artifacts, the dictionaries, lexicons, rules, and taxonomies.

If learning is supervised, the user should have an understanding on whether it is possible to improve the training data and thus the accuracy of the analysis. If the system is unsupervised, may the user provide training data or are the models "canned"? If the software uses active or reinforcement learning, how?

10 What is the resolution of automated text analysis?

Is the text analysis done at document, paragraph, sentence, phrase, or keyword-mention level? What features does the analysis extract – named entities, pattern-defined expressions, topics and themes, aspects (of an entity or topic), or relationships and attributes – and does it offer feature resolution, that is, identifying multiple features that are the same thing? (Winston Churchill, Mr. Churchill, the Prime Minister are a single individual.)

If the tool offers sentiment or emotions analysis, will it ascribe the sentiment/emotion to each of the resolved features or at some other level, and can the user choose the resolution of e.g. sentiment/emotion and semantic annotation?

11 Does the software provide document level data (e.g. individual posts to social media or specific survey open end) or only analytics based on document aggregation (i.e. quantitative analysis on a dashboard without the capability to drill through to the verbatims)?

In the social media context if the software does not provide drill-through access to the individual document, it will be difficult to check the accuracy of the analysis or to conduct any qualitative research type of analysis based on reading the actual verbatim.

12 In which languages can each of the automated analyses mentioned in questions 7-9 be carried out at the advertised accuracy?

This refers to text and audio and for text from online posts it includes not only natural languages such as English, French, Spanish, Mandarin etc. but also made-up languages such as Arab-ish or Greenglish (e.g., Arabic or Greek words written using Latin characters; here is an example from the Greek: χρόνος means time and in Greenglish is written like this: chronos or xronos).

13 Does the company use third party software or Web services (APIs) to produce the analyses or has it developed its own capability for market research purposes?

Numerous companies, ranging in size from small to large offer text analysis, sentiment, emotion, and other licensable libraries or Web services (APIs) for specific languages or image analysis. Whether applying language engineering or on supervised or unsupervised machine learning, the user should know if the company provides fully configured or training models or the end user will be responsible for that training in part or full. A custom or customizable model for a product category in a language will typically offer a higher accuracy.

14 Can the system extract or infer a data subject's demographic characteristics such as age, gender, income, education, and geography, and, if so, how (e.g. via metadata extraction, text analysis, or record linkage to external systems)? What validation processes are applied?

What processes does the company use to obtain this profile data, and how much of the data are populated? What proportions of the profile data are actual versus inferred versus unpopulated? How are inferred methods validated?

15 Is there any data sampling involved or needed, and if sampling is required or offered, what methods are applied?

Data sampling is relevant at the training data generation part of the process when the approach used is supervised machine learning. When harvesting from social media, sampling may help reduce the data cost albeit a small part of the overall cost, thus most users will use all the data available without any need for sampling.

16 What is the intended, target function of the system or service?

It is useful to know if the system was designed for key word monitoring, customer service, market research, or some other purpose. The main consideration for selecting a tool is the business objective so the service selected must suit the purpose.

Data quality and validation

17 How is the data cleaned to ensure that only relevant documents are used for the intended analysis?

Irrelevant comments or noise may result through homonyms, non-substantive side comments, spam, bots, etc. It is important for the user to know what measures are taken to ensure that data is clean and validated. The opposite problem to including irrelevant documents is not finding all the relevant documents in a larger body of text. Stemming or root words (e.g. by including the stem “wait“ to be able to harvest every word with any ending such as wait-list, wait-ed, wait-ing etc.) and including misspellings are some of the ways to make sure the number of missed documents is mimimised.

18 At the resolution mentioned in Q9 what is the minimum guaranteed accuracy of the analysis carried out by the software?

Accuracy typically is expressed in a number of different ways:

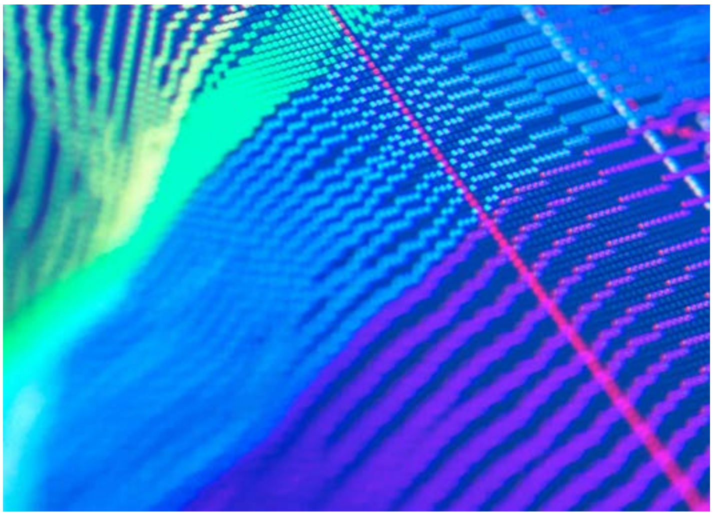
- Precision is the ratio of correctly predicted position observations to the total predicted positive observations, that is, the percent correct.
- Recall is the ratio of correctly predicted positive observations to all observations in the class.
- F1 Score is the weighted average of Precision and Recall.
- BLEU-1 is a measure of the similarity between how a model describes an image and how a human would describe it by looking at the words the descriptions have in common.

19 Is the user able to check the accuracy by themselves without any support from the software vendor?

The best way to verify the precision and recall is to extract a random sample of documents from the body of text and/or images and manually ascertain the level of agreement of a human judge with the software annotations.

20 What is the method for identifying spam in social media?

Biases might be created if spam is included in the data set. What types of automated and/or manual processes are in place for identifying spam or commercial messages that are misrepresented as consumer data? Are clients able to flag and identify such data and delete it themselves?



Ethical and legal compliance

21 Does the company comply with the relevant legal data protection requirements in the jurisdictions in which it sources, processes, and shares data?

Most types of unstructured data qualify as personal data, defined as “any information relating to a natural living person that can be used to identify an individual, for example by reference to direct identifiers (such as a name, specific geographic location, telephone number, picture, sound or video recording) or indirectly by reference to an individual's physical, physiological, mental, economic, cultural or social characteristics.” [1] Different countries and regions have different privacy laws governing how personal data must be protected. In general, companies are required to follow the local laws in all countries where they collect and process data, not just in the country where they are based or mainly operate. Users of the company services may also wish to ensure that the company assumes contractual liability for compliance with all applicable laws.

22 What specific processes are in place to ensure the above described compliance?

A company should regularly check the Terms of Use of all services (e.g. Facebook or Twitter) from which it collects data. When a company uses an API it still must have permission to collect or access the services' data. In the case where the company is provided with client-sourced data, it is important to understand the company's approach to data ownership, confidentiality, and protection.

23 What codes of conduct and industry standards does the company abide by?

There are many self-regulatory codes of conduct and industry standards including the ICC/ESOMAR International Code on Market, Opinion and Social Research and Data Analytics; the national research association codes of the 38 associations that comprise GRBN; ISO 20252 – Market, opinion, and social research; and other potentially relevant codes of conduct outside of the research industry.

[1] ICC/ESOMAR International Code on Market, Opinion, and Social Research and Data Analytics.

24 How does the company ensure that data subjects are not harmed as a direct result of their data being collected, processed, and shared?

Harm in this context is broadly defined as "tangible and material harm (such as physical injury or financial loss), intangible or moral harm (such as damage to reputation or goodwill), and excessive intrusion into private life, including unsolicited personally-targeted marketing messages." [2]

Researchers and subcontractors alike have a responsibly to ensure that use of the data will not result in harm to data subjects and there are measures in place to guard against such harm. Companies should have processes in place to assess systematically the potential for harm to data subjects and employ measures to mitigate against such harm. Examples of harm include use of data in ways that discriminate against individuals in decisions regarding extension of credit, considerations of employment, or denial of medical care, to name a few. One especially prominent use of social media postings is identification of individuals to receive personally targeted marketing messages.

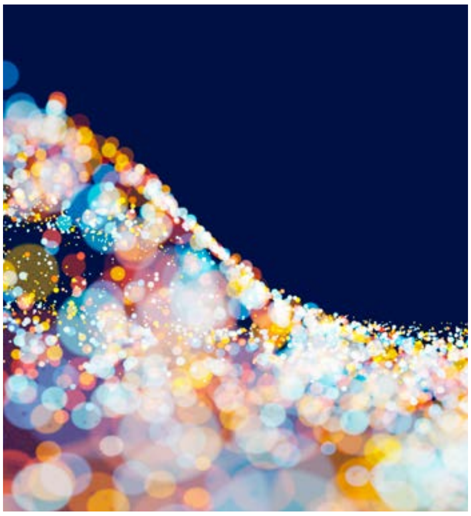
25 How does the company safeguard the privacy of data subjects in what it shares with users?

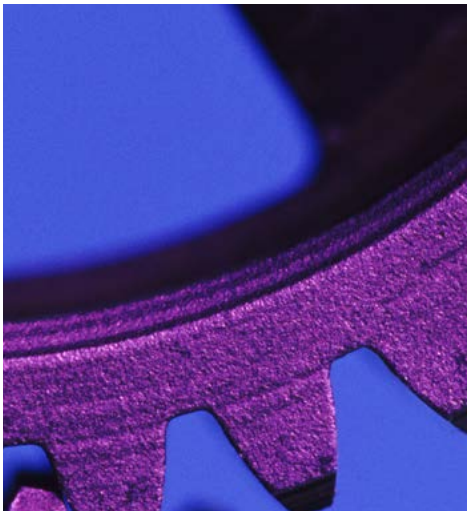
Unstructured data can be collected from many sources, both public and private. Examples of public sources where data may be considered to be in the public domain include Twitter, news sites, Reviews, and most blogs. Examples of private sources include so-called "walled gardens" social media services (e.g. Facebook), research communities, surveys, call center logs, etc. Some companies may distinguish between public and private data in what they share with their customers, sharing raw data in the case of the former and masking data to protect data subject identities in the case of the latter. Or, they may make no distinctions, relying instead on the customer to protect data subjects' identities. In any case, the criteria governing what is shared and in what form is an important consideration. See also Question 5.

26 What information security practices are in place to ensure the security of data? Does the company allow clients to audit said processes?

Companies should have well-documented information security practices to ensure that personal data is thoroughly protected from inadvertent disclosure or loss. Review of those practices is especially important when you plan to provide your own data for processing and analysis.

[2] Ibid.





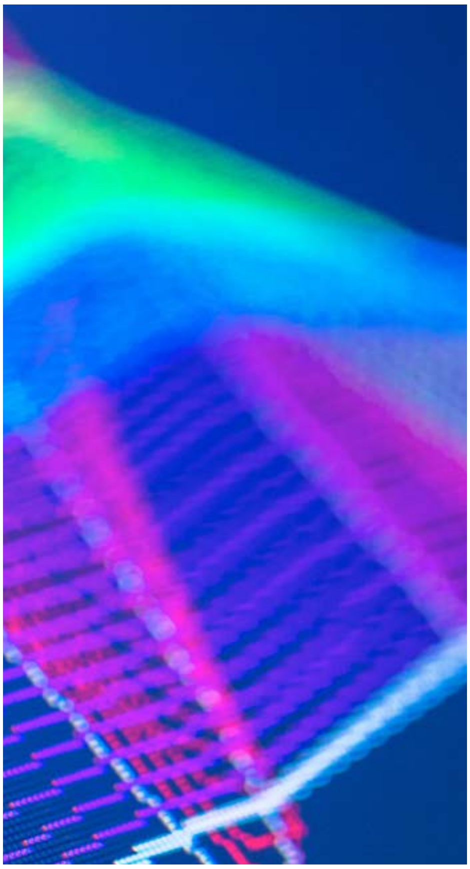
Guidance on professional standards

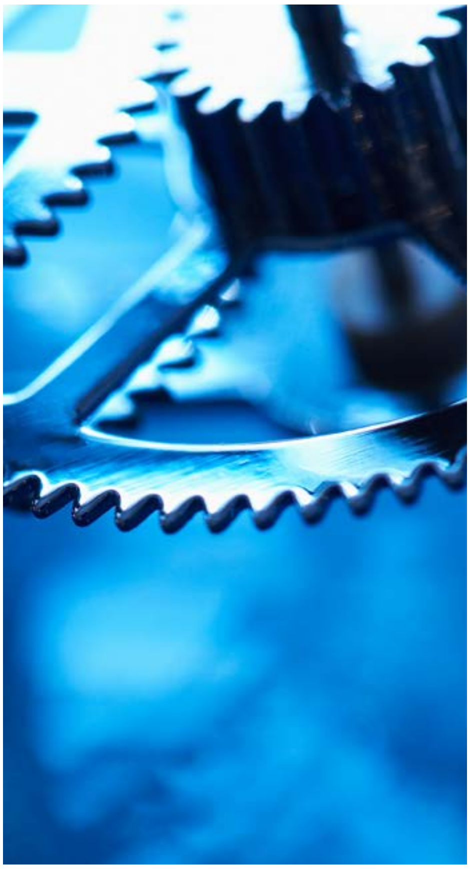
Maintaining consumer trust is integral to effective market, social and opinion research. ESOMAR through its codes and guidelines promotes the highest ethical and professional standards for researchers around the world. Providers of social media research are expected to follow all relevant ESOMAR codes and guidelines.

The ICC/ESOMAR International Code on Market, Opinion, and Social Research and Data Analytics, which was developed jointly with the International Chamber of Commerce, sets out global fundamentals for self-regulation for researchers. It has been undersigned by all ESOMAR members and adopted or endorsed by more than 60 national market research associations worldwide.

The ESOMAR Guideline on Social Media Research is of particular relevance to researchers using social media data and should be read in conjunction with these questions for more explanation of the legal and professional responsibilities of researchers who are collecting and analyzing social media data.

For further guidance visit the Codes and Standards section of the ESOMAR website or contact professional.standards@esomar.org.





Project team

Co-chairs

Michalis Michael, CEO DigitalMR
Reg Baker, Consultant to ESOMAR Professional Standards Committee

Members

Joaquim Bretcha, ESOMAR Council Member and International Director, NetQuest
Francesco D'Orazio, Chief Innovation Officer, Pulsar
Seth Grimes, Consultant Alta Plana Corporation
Jan Kringels, Head of Global Net Promoter Programme Vodafone UK
Michelle Roseman Turner, MaritzCX
Annelies Verhaeghe, Managing Partner, Insites Consulting

ESOMAR
Atlas Arena, Azië building
Hoogoorddreef 5
1101 BA Amsterdam
The Netherlands
Tel: +31 20 664 2141
www.esomar.org