ESTUDOS AVANÇADOS EM CIÊNCIA DA INFORMAÇÃO: PESQUISA EM METADADOS E WEBSEMÂNTICA

Aula 05 – Das hierarquias as redes: taxonomias, tesauros e ontologias

Prof. Dalton Martins

daltonmartins@unb.br

2 sem 2019

UnB

UnB, sua linda

meu orgulho é você

Relembrando...

Term Lists

Authority Files

Authority Files are lists of terms that are used to control the variant names for an entity or the domain value for a particular field. Examples include names for countries, individuals, and organizations. Non-preferred terms may be linked to the preferred versions. This type of KOS generally does not include a deep organization or complex structure. The presentation may be alphabetical or organized by a shallow classification scheme. There may be some limited hierarchy applied in order to allow for simple navigation, particularly when the authority file is being accessed manually or is extremely large. Specific examples of authority files include the Library of Congress Name Authority File and the Getty Geographic Authority File.

Glossaries

A glossary is a list of terms, usually with definitions. The terms may be from a specific subject field or those used in a particular work. The terms are defined within that specific environment and rarely have variant meanings provided. Examples include the EPA Terms of the Environment.

Gazetteers

A gazetteer is a dictionary of place names. Traditional gazetteers have been published as books or they appear as indexes to atlases. Each entry may also be identified by feature type, such as river, city, or school. Geospatially referenced gazetteers provide coordinates for locating the place on the earth's surface. An example is the Geographic Names Information Service <http://www-nmd.usgs.gov/www/gnis/>. Note that the term "gazetteer" has several other meanings including an announcement publication such as a patent or legal gazetteer. These gazetteers are often organized using classification schemes or subject categories.

Dictionaries

Dictionaries are alphabetical lists of terms and their definitions that provide variant senses for each term, where applicable. They are more general in scope than a glossary. They may also provide information about the origin of the term, variants (both by spelling and morphology), and multiple meanings across disciplines. While a dictionary may also provide synonyms and through the definitions, related terms, there is no explicit hierarchical structure or attempt to group terms by concept.

https://nkos.slis.kent.edu/KOS_taxonomy.htm

Classification and Categorization

Subject Headings

This scheme provides a set of controlled terms to represent the subjects of items in a collection. Subject heading lists can be extensive, covering a broad range of subjects. However, the subject heading list's structure is generally very shallow, with a limited hierarchical structure. In use, subject headings tend to be pre-coordinated, with rules for how subject headings can be joined to provide more specific concepts. Examples include the Medical Subject Headings (MeSH) and the Library of Congress Subject Headings (LCSH).

Classification Schemes, Taxonomies and Categorization Schemes

These terms are often used interchangeably. Though there may be subtle differences from example to example, in general these types of KOSs provide ways to separate entities into "buckets" or relatively broad topic levels. Some examples provide a hierarchical arrangement of numeric or alphabetic notation to represent broad topics. These types of KOSs may not follow the strict rules for hierarchy required in the ANSI NISO Thesaurus Standard (Z39.19) (NISO), and they lack the explicit relationships presented in a thesaurus. Examples of classification schemes include the Library of Congress Classification Schedules (an open, expandable system), the Dewey Decimal Classification (a closed system of 10 numeric sections with decimal extensions), and the Universal Decimal Classification (based on Dewey but extended to include facets). Subject categories are often used to group thesaurus terms in broad topic sets, outside the hierarchical scheme of the thesaurus. Taxonomies are increasingly being used in object oriented design and knowledge management systems to indicate any grouping of objects based on a particular characteristic. "Taxonomy" may also refer to a scheme that presents biota in a hierarchical arrangement based on some characteristic.

https://nkos.slis.kent.edu/KOS_taxonomy.htm

Relationship Groups

Thesauri

These KOSs are based on concepts, and they show relationships between terms. Relationships commonly expressed in a thesaurus include hierarchy, equivalence, and associative (or related). These relationships are generally represented by the notation BT (broader term), NT (narrower term), SY (synonym), and RT (associative or related). Associative Relationships may be more granular in some schemes. For example, the Unified Medical Language System (UMLS) from the National Library of Medicine has defined over 40 relationships, many of which are associative in nature. Preferred terms for indexing and retrieval are identified. Entry terms (or non-preferred terms) point to the preferred terms that are to be used for each concept.

There are standards for the development of monolingual thesauri (NISO, 1998; ISO, 1986) and multi-lingual thesauri (ISO, 1985). However, in these standards the definition of a thesaurus is fairly narrow. Standard relationships are assumed, as well as the identification of preferred terms, and there are specific rules for the creation of the relationships between terms. It should be noted that the definition of a thesaurus in these standards is often at variance with schemes that are actually called thesauri. There are many thesauri that do not follow all the rules of the standard, but are still generally thought of as thesauri. Note: Another type of "thesaurus" represents only equivalence (synonymy), such as the Roget's Thesaurus (with the addition of classification categories).

Many thesauri are very large (more than 50,000 terms). Most were developed for a specific discipline, or to support a specific product or family of products. Examples include the Food and Agricultural Organization's Aquatic Sciences and Fisheries Thesaurus and the NASA Thesaurus for aeronautics and aerospace-related topics.

Semantic Networks

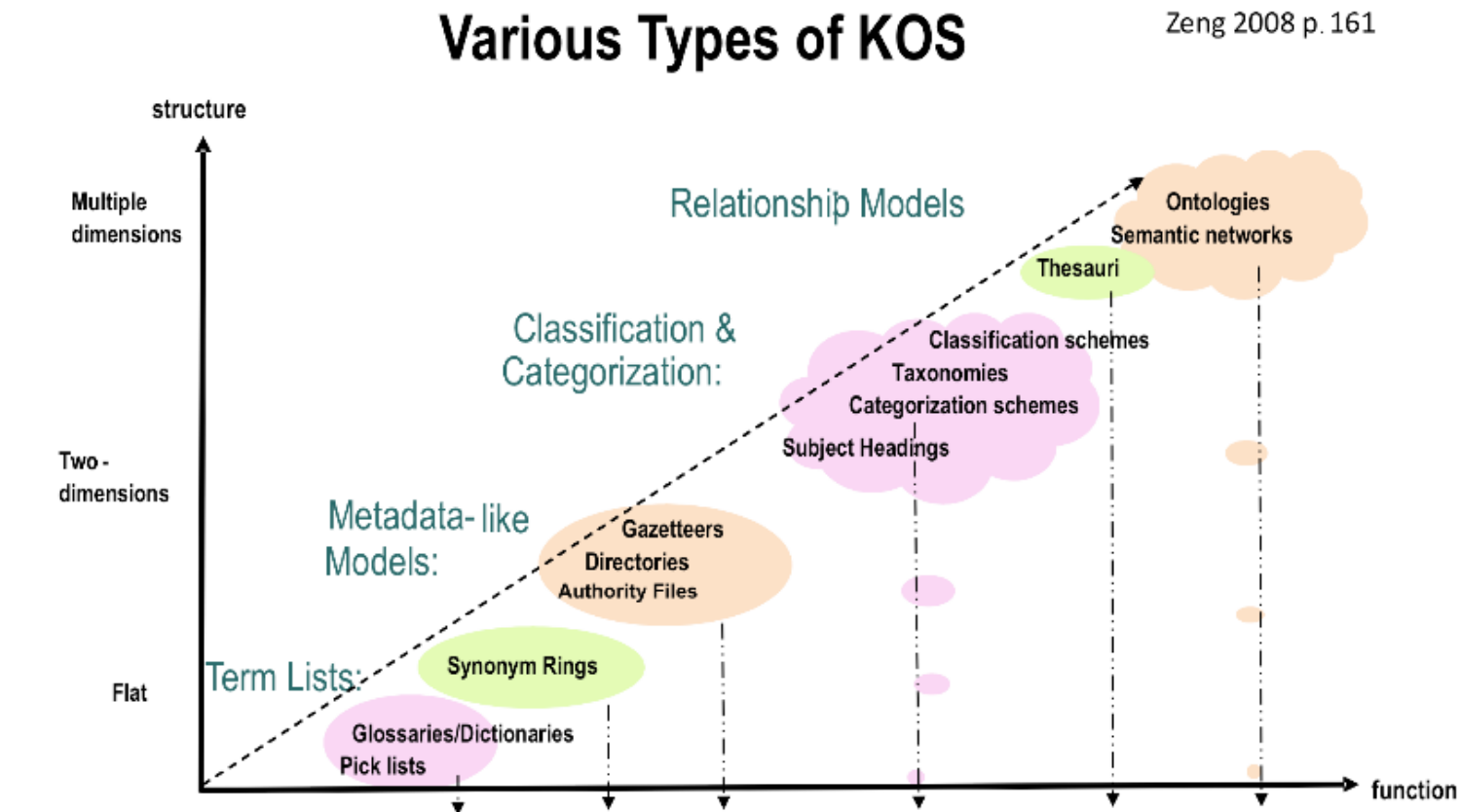
With the advent of natural language processing, there have been significant developments in the area of semantic networks. These KOSs structure concepts and terms not as hierarchies but as a network or a Web. Concepts are thought of as nodes with various relationships branching out from them. The relationships generally go beyond the standard BT, NT and RT. They may include specific whole-part relationships, cause-effect, parent-child, etc. One of the most noted semantic network is Princeton's WordNet, which is now used in a variety of search engines.

Ontologies

"Ontology" is the newest label attached to some KOSs. Ontologies are being developed as specific concept models by the Knowledge Management community. They can represent complex relationships between objects, and include the rules and axioms missing from semantic networks. Ontologies that describe knowledge in a specific area are often connected with systems for data mining and knowledge management.

https://nkos.slis.kent.edu/KOS_taxonomy.htm

Figure 1. Various Types of KOS(2)



Major functions

eliminating ambiguity	xxx		xxx	xx	xxxx	xx
controlling synonyms		xxxx	xxx	xx	xxxx	xx
establishing relationships: hierarchical			x	xxxx	xxx	xxx
establishing relationships: associative					xxxx	xxxxx
presenting properties						xxxxx

https://nkos.slis.kent.edu/KOS_taxonomy.htm

Knowledge Organization Systems (KOS)†

Marcia Lei Zeng

Kent State University, School of Library and Information Science,
Kent, OH, USA 44242-0001, <mzeng@kent.edu>

Marcia Lei Zeng has been involved in the development, teaching, and research of knowledge organization systems (KOS) for over 20 years. She has served on standards committees and working groups for IFLA, Special Libraries Association (SLA), American Society for Information Science and Technology (ASIST), and US National Information Standards Organization (NISO). She is a member of the Advisory Group for NISO Z39.19-2005 for monolingual controlled vocabularies. Her services include chairs of the SLA Technical Standards Committee, ASIST Standards Committee, IFLA Classification and Indexing Section, and IFLA Functional Requirements for Subject Authority Records (FRSAR) Working Group.

† The author would like to thank the following publishers of vocabularies, software, and websites that were used in the examples of this paper: NISO Press, the National Library of Medicine, Google, OCLC, University of California Santa Barbara, University of Arizona, Open Directory Project, Kent State University, J. Paul Getty Trust, Drexel University, University of Glamorgan, University of Washington, and the Gene Ontology Consortium. Permission to reprint copyrighted material was granted from: NISO, Denise Bedford, Karl Fast, Tree of Life Web Project, Maja Zumer, Vocabulary Program of the J. Paul Getty Trust, Xia Lin, Douglas Tudhope, Foundational Model of Anatomy Ontology, and the Gene Ontology Consortium.

Zeng, Marcia Lei. **Knowledge Organization Systems (KOS).** *Knowledge Organization, 35(3/2)*, 160-182. 39 references.

ABSTRACT: Knowledge organization systems (KOS) can be described based on their structures (from flat to multidimensional) and main functions. The latter include eliminating ambiguity, controlling synonyms or equivalents, establishing explicit semantic relationships such as hierarchical and associative relationships, and presenting both relationships and properties of concepts in the knowledge models. Examples of KOS include lists, authority files, gazetteers, synonym rings, taxonomies and classification schemes, thesauri, and ontologies. These systems model the underlying semantic structure of a domain and provide semantics, navigation, and translation through labels, definitions, typing, relationships, and properties for concepts.

The term knowledge organization systems (KOS) is intended to encompass all types of schemes for organizing information and promoting knowledge management, such as classification schemes, gazetteers, lexical databases, taxonomies, thesauri, and ontologies (Hodge 2000). These systems model the underlying semantic structure of a domain and provide semantics, navigation, and translation through labels, definitions, typing, relationships, and properties for concepts (Hill et al. 2002, Koch and Tudhope 2004). Embodied as (Web) services, they facilitate resource discovery and retrieval by acting as semantic road maps, thereby making possible a common orientation for indexers and future users, either human or machine (Koch and Tudhope 2003, 2004).

https://www.researchgate.net/publication/297530633_Knowledge_Organization_Systems_KOS

Reflexões sobre os principais problemas relacionados a dificuldade do uso de instrumentos de bons instrumentos de gestão da informação e como isso se relaciona com metadados, taxonomias, tesauros e ontologias....

http://www.kmworld.com/Articles/Editorial/Features/The-high-cost-of-not-finding-information-9534.aspx

Studies by IDC, as well as organizations such as the Working Council of CIOs, AIIM, the Ford Motor Company and Reuters have found that:

- Knowledge workers spend from 15% to 35% of their time searching for information.
- Searchers are successful in finding what they seek 50% of the time or less, according to both Web search engines and our own surveys. An IDC study in 2001 ("Quantifying Enterprise Search," IDC, May 2002) found that only 21% of respondents said they found the information they needed 85% to100% of the time (see Figure 1).
- 40% of corporate users reported that they can not find the information they need to do their jobs on their intranets.

http://www.kmworld.com/Articles/Editorial/Features/The-high-cost-of-not-finding-information-9534.aspx

How much time is spent reworking or recreating information because it has not been located?

Recent research on knowledge work shows that knowledge workers spend more time recreating existing information than they do turning out information that does not already exist. Some studies suggest that 90% of the time that knowledge workers spend in creating new reports or other products is spent in recreating information that already exists. In 1999, a European study by IDC examined that phenomenon, called the "knowledge work deficit," and concluded that the cost of intellectual rework, substandard performance and inability to find knowledge resources was $5,000 per worker per year.

Using those studies as a basis, we set out to quantify the impact that not finding information might have on a typical enterprise of a thousand knowledge workers who earned an average salary plus benefits of $80,000 a year. We looked at:

http://www.kmworld.com/Articles/Editorial/Features/The-high-cost-of-not-finding-information-9534.aspx

Here's what we found

- The time spent looking for and not finding information costs our mythical organization a total of $6 million a year. That doesn't include opportunity costs or the costs of reworking information that exists but can't be located.
- The cost of reworking information because it hasn't been found costs that organization a further $12 million a year (15% of time spent in duplicating existing information).
- Not locating and retrieving information has an opportunity cost of more than $15 million annually. Accelerating the introduction of a blockbuster drug or delaying its demotion to generic status by just one day through use of information access software could mean $8.5 million or more each day.
- Increased e-commerce revenue pays for the improved search software in a couple of months. Companies like Charles Schwab, Lands' End, Staples or Macy's have increased their commerce revenue by amounts like $125,000 per month, or 400% in average deal size.
- Call center costs and volumes have been decreased by 30% and more when better search and browsing tools were implemented.

http://www.kmworld.com/Articles/Editorial/Features/The-high-cost-of-not-finding-information-9534.aspx

FEATURE

Have You Ever Spent 38 Minutes Looking for a Document?

In the age of Google search, too many documents and pieces of data are still hiding inside company networks, a new survey says.

By Thomas Wailgum

CIO | MAY 13, 2009 7:00 AM PST

In this age of sub-one-second Google search queries, modern networking and T1 Internet connections, it seems ludicrous that any type of information-based search would take even 38 *seconds* to complete.

MORE ON CIO.com

The Ugly Truth About One Version of the Truth

[Beware the 9 warning signs of bad IT architecture and see why these 10 old-school IT principles still rule. | Sign up for CIO newsletters.]

Data Management Danger: Less Than Half of MDM Plans Are Effective

Companies Struggle to Find the Truth in Massive Data Flows

https://www.cio.com/article/2428183/have-you-ever-spent-38-minutes-looking-for-a-document-.html

A new survey, however, finds that employees at big companies (with more than 10,000 employees) spend, on average, 38 *minutes* searching for one document -- whether that's on their own computers or their organization's networks, databases or intranet.

In addition, the survey of 200 respondents from companies in a wide range of industries found that employees are having trouble finding the most efficient and appropriate technology tools to locate documents or internal expertise. To find in-house experts, for instance, 71 percent of the respondents said they "ask around"; 46 percent said they use the company directory; 34 percent use the company website or intranet; and 30 percent said they send a companywide e-mail (and we all know how annoying those can be).

https://www.cio.com/article/2428183/have-you-ever-spent-38-minutes-looking-for-a-document-.html

McKinsey&Company

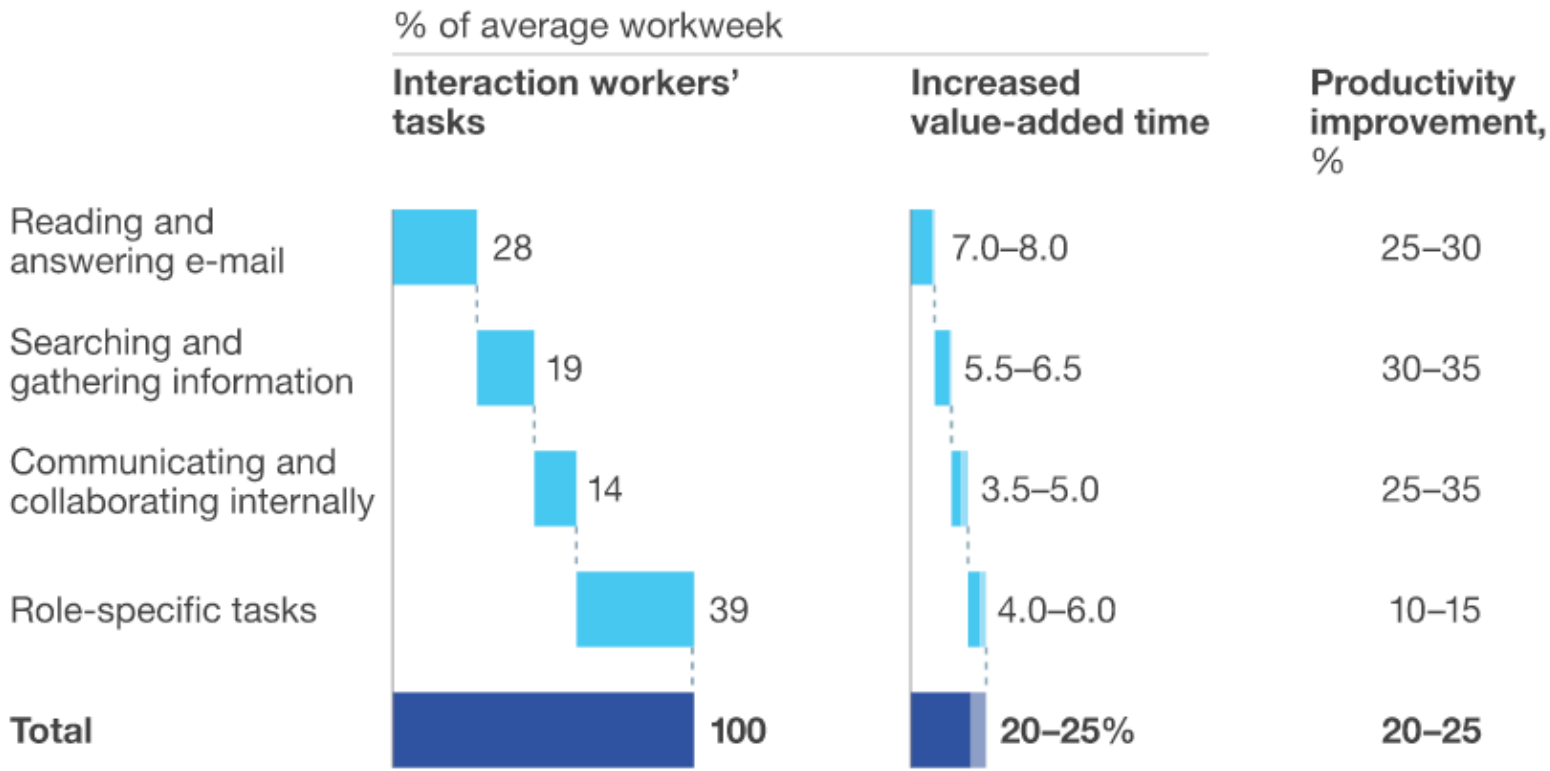
McKinsey Global Institute

July 2012

The social economy: Unlocking value and productivity through social technologies

https://www.mckinsey.com/~/media/McKinsey/Industries/Technology%20Media%20and%20Telecommunications/High%20Tech/Our%20Insights/The%20social%20economy/MGI_The_social_economy_Executive_Summary.ashx

Improved communication and collaboration through social technologies could raise the productivity of interaction workers by 20 to 25 percent.



Source: International Data Corporation (IDC); McKinsey Global Institute analysis

https://www.mckinsey.com/~/media/McKinsey/Industries/Technology%20Media%20and%20Telecommunications/High%20Tech/Our%20Insights/The%20social%20economy/MGI_The_social_economy_Executive_Summary.ashx

European Commission

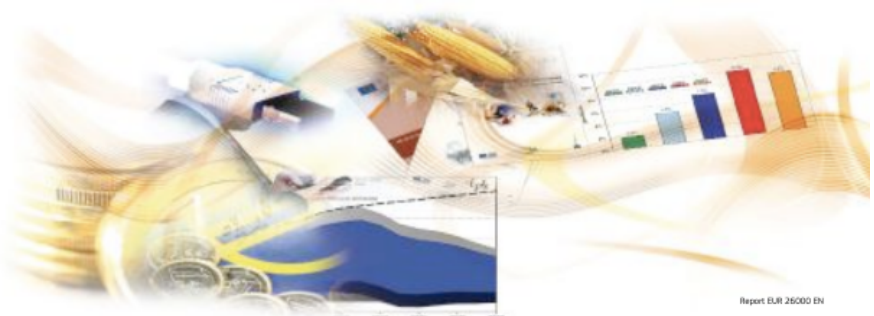
JRC SCIENTIFIC AND POLICY REPORTS

Enterprise Search in the European Union: A Techno-economic Analysis

Authors: Martin White, Stavri G Nikolov

Editors: Shara Monteleone, Ramon Compaño, Ioannis Maghiros

2013



Report EUR 26000 EN

Joint Research Centre

http://publications.jrc.ec.europa.eu/repository/bitstream/JRC78202/jrc78202.pdf

Motivators

Motivators for the development of an enterprise search market, as emerged from the surveys mentioned in this report and also from the workshop organized by JRC-IPTS, "Exploring the future of Enterprise Search", in Seville in October 2011 are:

- There is increasing information everywhere: more than 200 billion emails per day; 80% of enterprise information is unstructured.
- Digital data growth is enormous: it is expected to be 35 zettabytes in 10 years' time. In particular, it seems that 94% of organizations are collecting and managing more business data than just a few years ago and business information collected/managed has increased by 86% in the last few years.[1]
- The cost of poor data management: organizations are seemingly losing revenue each year (on average, 14%) as a result of not being able to fully leverage the information they collect. That translates to circa $130 million in lost opportunity each year for a $1 billion organization.[2]
- Legal compliance of the enterprise: obligation to store and find all enterprise documents, business communications for legal reasons. Enterprise data is all over the place. ES has to federate all the information existing in both structured data (databases) and unstructured data (text, reports, mail).

http://publications.jrc.ec.europa.eu/repository/bitstream/JRC78202/jrc78202.pdf

Although the issue is acknowledged to be of extreme importance, only a small number of companies benefit from provision of dedicated enterprise search technologies and products. The global enterprise search business probably has no more than 200 companies. There are around **60 enterprise search vendors** who together account for probably over 90% of enterprise search software sales, excluding open source products where there is no license fee. **Six** vendors, all of them multinational IT companies, have a major impact on the development of search technology but only a limited impact on the development of the enterprise search market through promotional activities. These can be termed Type 1 vendors, and are Google, HP, Lexmark, Microsoft, IBM and Oracle. **Type 1 companies** in general sell enterprise search as part of an overall enterprise application suite, and not as stand-alone products. The long-term implications of the acquisition of Autonomy by HP in 2011 will not be clear for some time.

The other companies form **a large Type 2 category**. The primary characteristic of Type 2 companies is that they have developed proprietary search software and are mostly funded by venture capital and private equity investments. Because of their small size and the fact that they are private companies, there are no requirements to publish detailed accounts of revenues and costs. Most of these companies have revenues of less than $20 million. Examples of Type 2 companies with headquarters in the EU are Fabasoft (Austria) and Sinqeua (France).

Type 3 companies build products around open-source software such as Lucene/Solr, with Intrafind (Germany) being an example.

http://publications.jrc.ec.europa.eu/repository/bitstream/JRC78202/jrc78202.pdf

Specifying and selecting search software

One main barrier to making a business case is a **lack of awareness of the functionality** of enterprise search applications and the benefits that effective search can have for the enterprise. Enterprise search is of potential value to most, if not all, employees, but no single department wishes to take responsibility for making a business case. There is evidence that implementing enterprise search is not a high priority; the reason being that one single business unit is probably unable to make a business case for enterprise search. Organisations have not usually done any research into the most important tasks carried out by employees and the extent to which enterprise search would improve operational effectiveness.

IT departments have a role to play in the technical evaluation of enterprise search applications and in the initial installation, but it is important that **business requirements** are well defined in developing a business case for selecting or replacing an enterprise search application. IT managers, however, may have little formal teaching about information retrieval technologies and may not be aware of how to evaluate search applications or how to plan adequately for implementation and subsequent optimisation.

http://publications.jrc.ec.europa.eu/repository/bitstream/JRC78202/jrc78202.pdf

For a large enterprise search implementation there are a number of roles that need to be filled. These are a Search Manager, a Search Technology Manager, an Information Specialist with a strong business background, a Search Analytics Manager and a Search Support Manager. For all these posts the team need to have a good background in the technology and implementation of enterprise search. However, the EU seems to have a significant lack of academic institutions that are offering taught courses in information retrieval. There are around 30 institutions in the EU undertaking research into enterprise search applications, but there are no full-time three year undergraduate courses. As a result there is a **shortage of skilled professionals** to join search vendors as development and implementation engineers, and to join enterprise search support teams.

http://publications.jrc.ec.europa.eu/repository/bitstream/JRC78202/jrc78202.pdf

Technology forecast

The outcomes of the Delphi study, carried out by IPTS in collaboration with Intranet Focus Ltd in 2011, indicates **six important areas** of technical development for enterprise search over the next few years, namely

- Integrated search of structured and unstructured content;
- Search as an integration platform (unified access platforms);
- Search incorporated into business intelligence applications;
- Search-based applications;
- Text mining, and
- Enterprise mobile search applications.

http://publications.jrc.ec.europa.eu/repository/bitstream/JRC78202/jrc78202.pdf

Future trends in ES: cloud-based and user-demand approach, open data models, interoperability

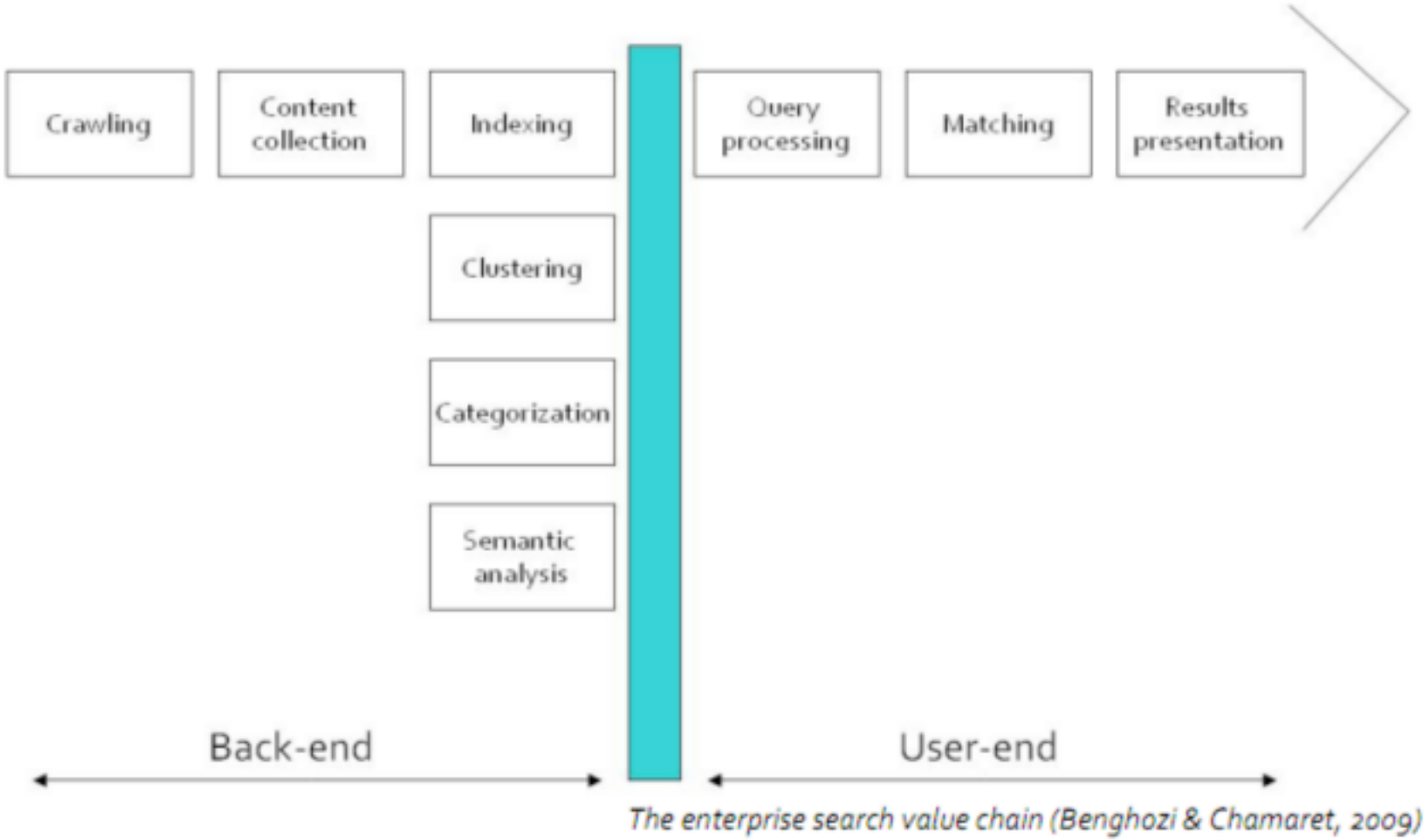
According to our Delphi-type study and to the results of the workshop on Enterprise Search (organized in Seville in October 2011), the technologies that are regarded as most important to the demand and adoption of ES are **Search-Based Applications (SBA)** and **integrated search platform (unified access platform)**. Integrated platform and search-based applications are key solutions because they promise to provide semantic linking (combining structured and unstructured data) and semantic search (allowing intelligent analysis of query).

http://publications.jrc.ec.europa.eu/repository/bitstream/JRC78202/jrc78202.pdf

To sum up:

- There is no dominant supplier of search applications in the EU.
- Although there are no reliable data nor are there any revenue analyses for the EU enterprise search market, we can assume that the market potential is still considerable, particularly for medium-sized companies that could benefit from using ES solutions.
- The total revenues of EU-headquartered search vendors are between €100 million and €200 million.
- Some of the main challenges that the ES market has to face are: how to define a stable and open architecture over which a European ecosystem could develop and mature, how component technology testing could be performed (see benchmark section), what would facilitate technology transfer to industry (see tech transfer section), or how to remedy the shortage of skilled professionals.
- A general issue to be addressed in the future is the fact that boundaries between enterprise search, text and data mining, business intelligence and content analytics are becoming very blurred.[4]
- To our knowledge, this report provides the most detailed and comprehensive techno-economic analysis of the enterprise search market in the EU and includes an up-to-date list of ES vendors.

http://publications.jrc.ec.europa.eu/repository/bitstream/JRC78202/jrc78202.pdf



The enterprise search value chain (Benghozi & Chamaret, 2009)

http://publications.jrc.ec.europa.eu/repository/bitstream/JRC78202/jrc78202.pdf

Search and Discovery
- exploiting knowledge, minimizing risk

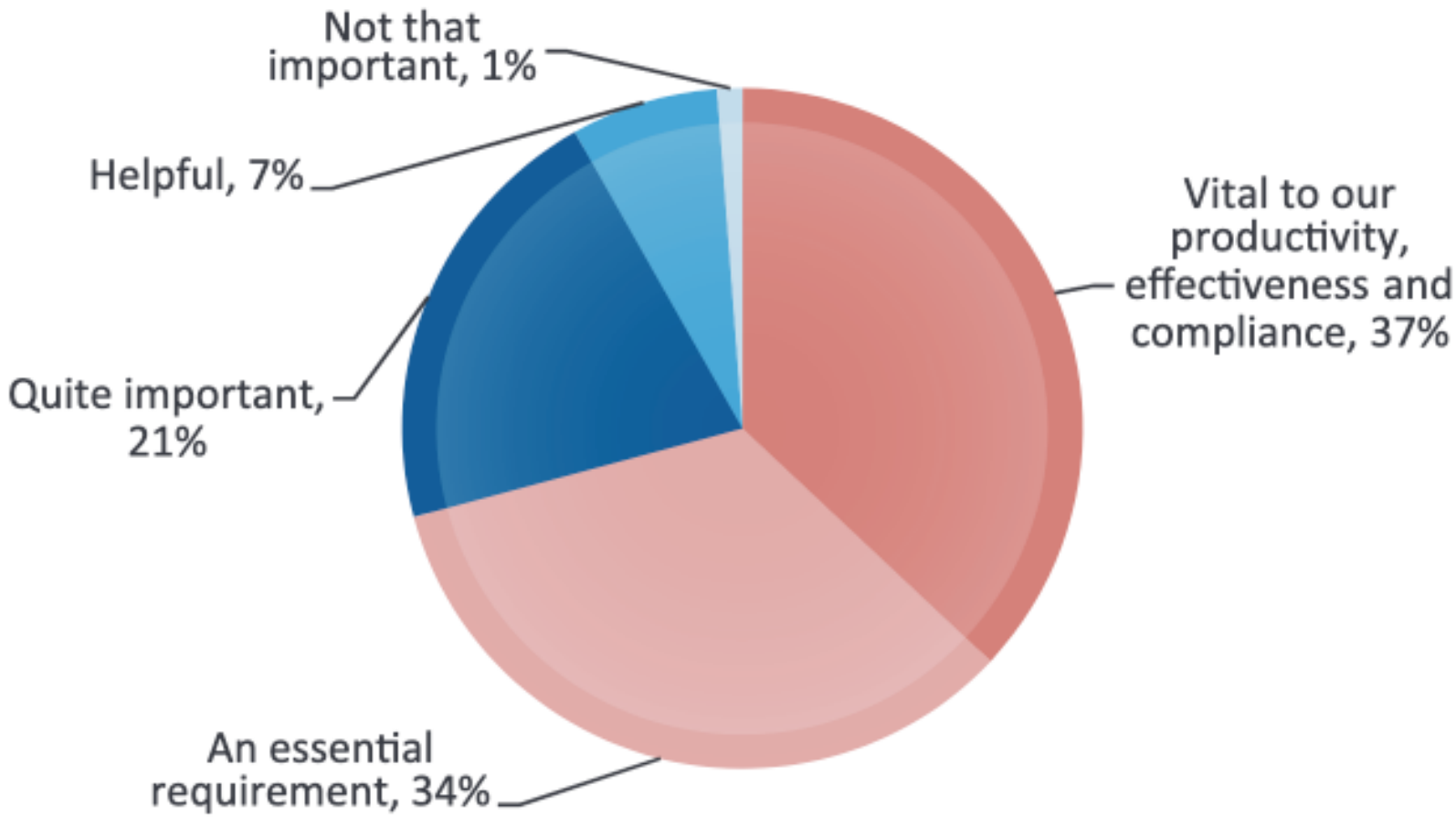
Underwritten in part by:

BA INSIGHT · IRON MOUNTAIN · OPENTEXT · SEARCH · ZAPPROVED

aiim®
The Global Community of Information Professionals
aiim.org | 301.587.8202

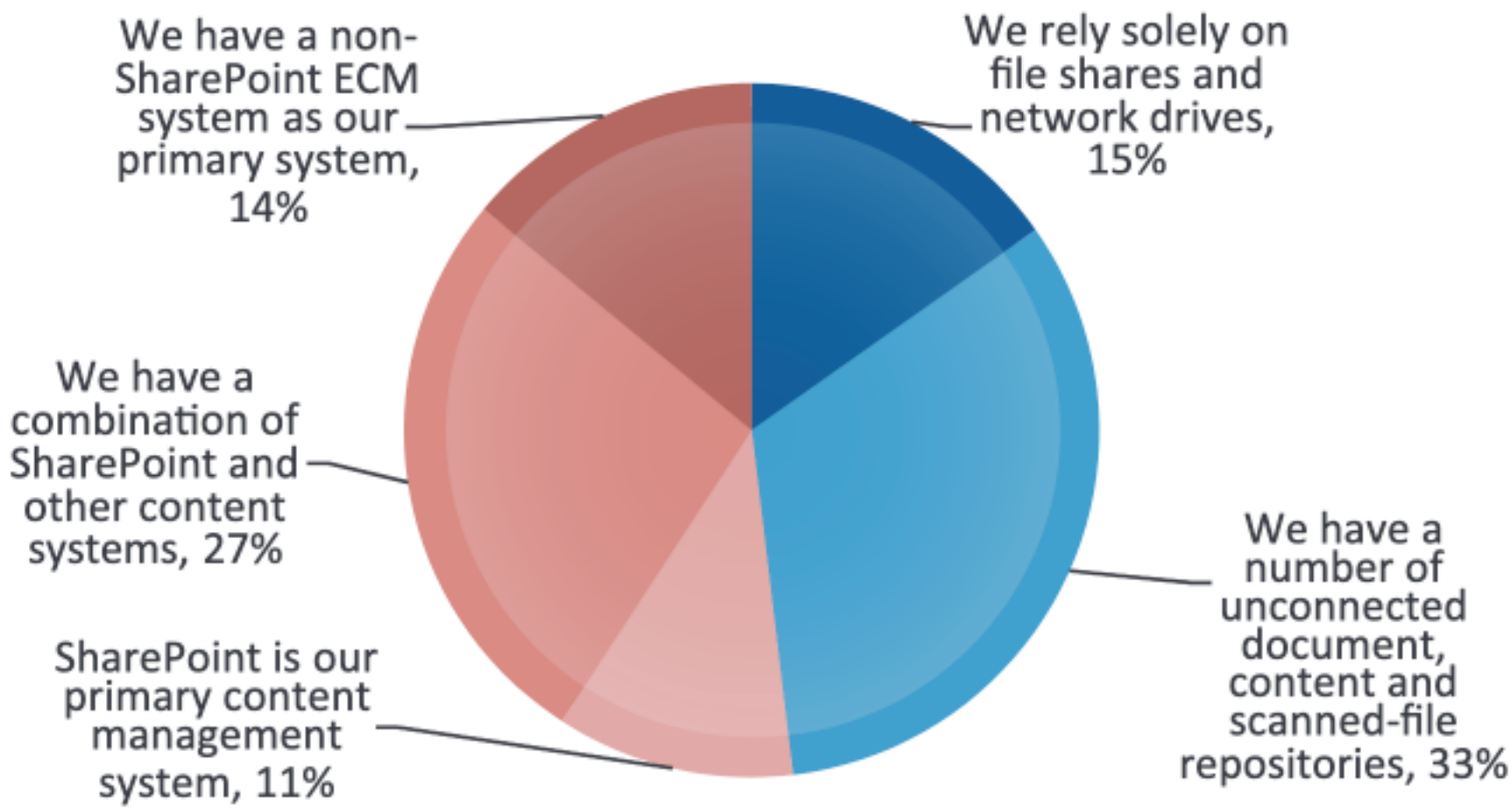
https://info.aiim.org/search-and-discovery

Figure 1: How important is it in your organization for employees to have an effective way to search internal content and documents in order to carry out their tasks? *(N=351)*



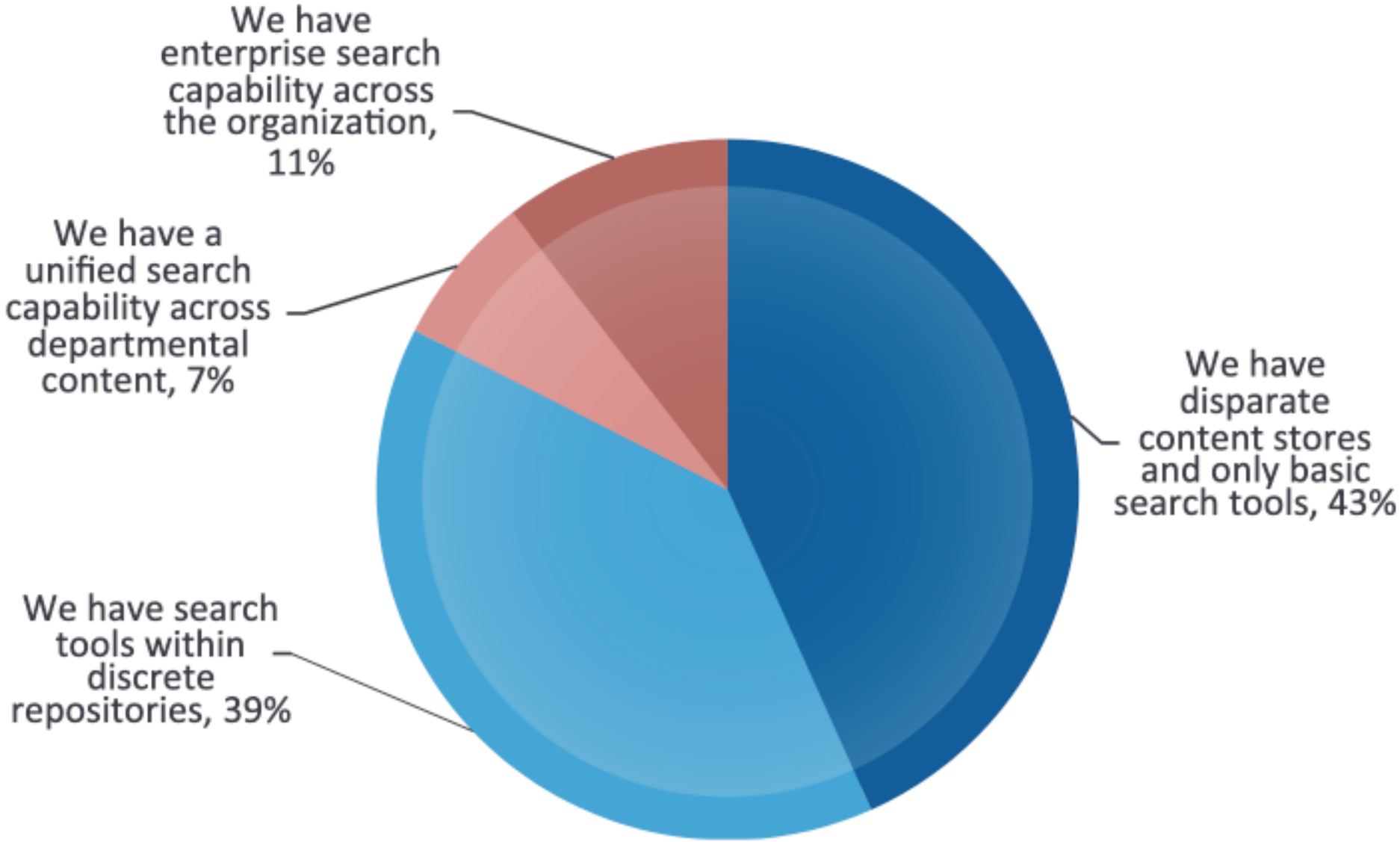
https://info.aiim.org/search-and-discovery

Figure 2: How would you describe your current enterprise content management (ECM) system(s)? (N=253)



https://info.aiim.org/search-and-discovery

Figure 3: How good is your ability to search across your key content? (Pick highest capability) *(N=350)*

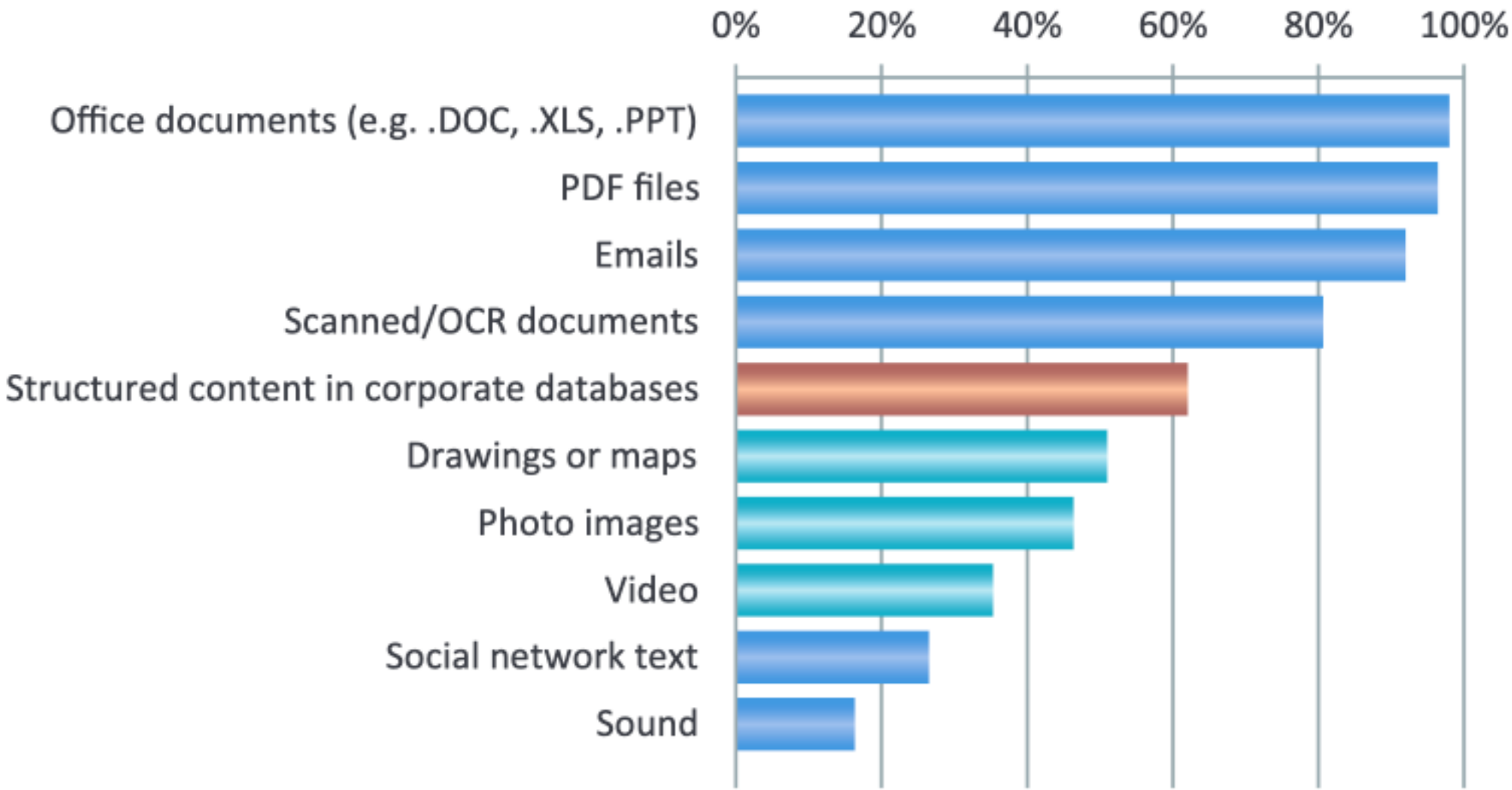


https://info.aiim.org/search-and-discovery

Figure 4: Thinking of the maturity of your approach to search, which of the following do you have?
(N=266 multiple)

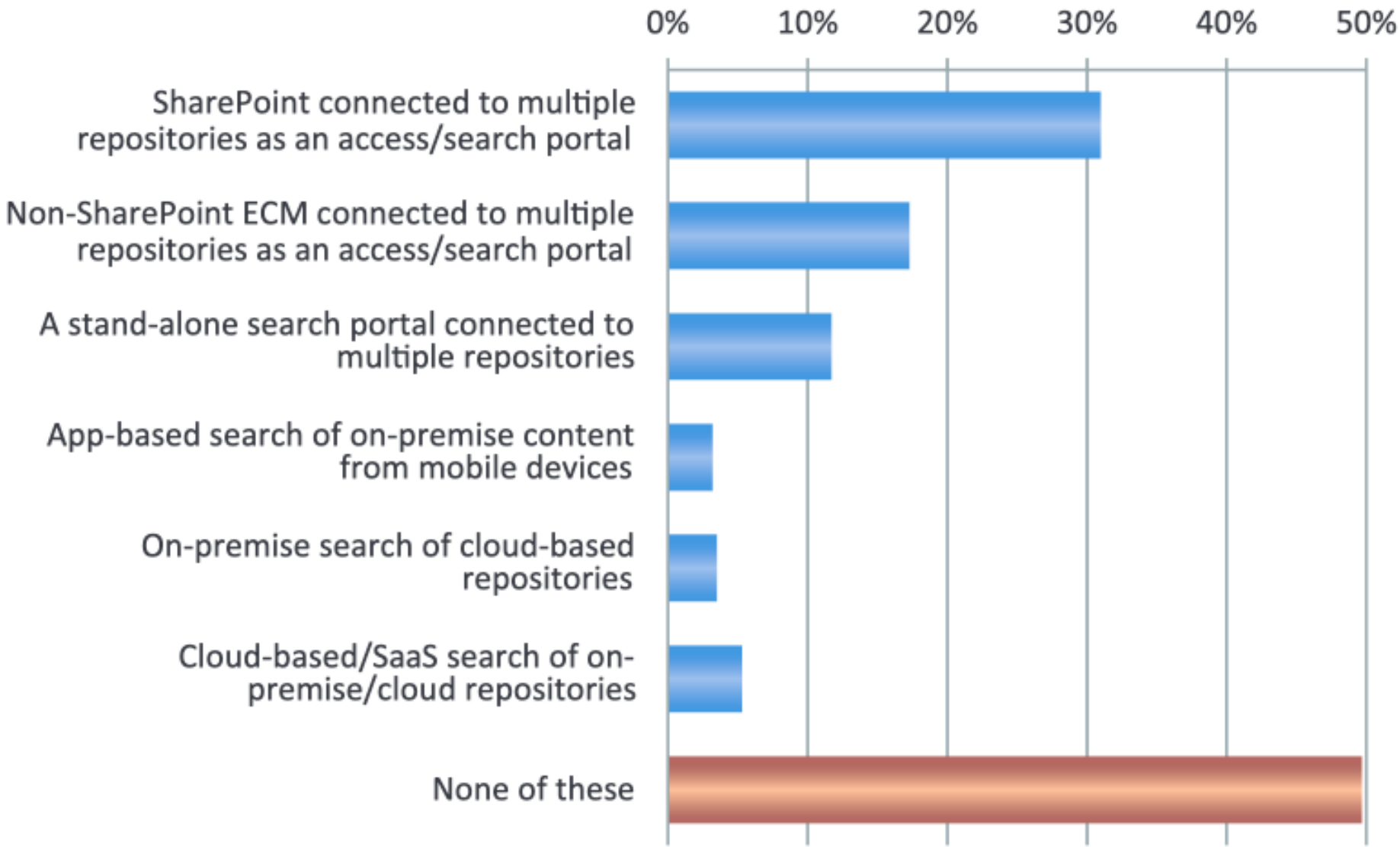
https://info.aiim.org/search-and-discovery

Figure 7: Which of the following content types is it important for your employees to be able to search? (Check those that are important). *(N=306)*



https://info.aiim.org/search-and-discovery

Figure 10: Do you have any of the following? *(N=342, multiple)*



https://info.aiim.org/search-and-discovery

ENTERPRISE SEARCH
AND FINDABILITY
SURVEY 2016

FINDWISE

https://findwise.com/en/Enterprise-Search-Findability-Report-2016

»Findability dimensions and success factors«

Business | Users | Information | Organisation | Technology

BUSINESS

- Search and Findability strategy
- Involving business stakeholders
- Key Performance Indicators

USERS

- User target groups
- Tailored search experiences
- User feedback process

INFORMATION

- Taxonomy
- Metadata standard
- Content life cycle management process

ORGANISATION

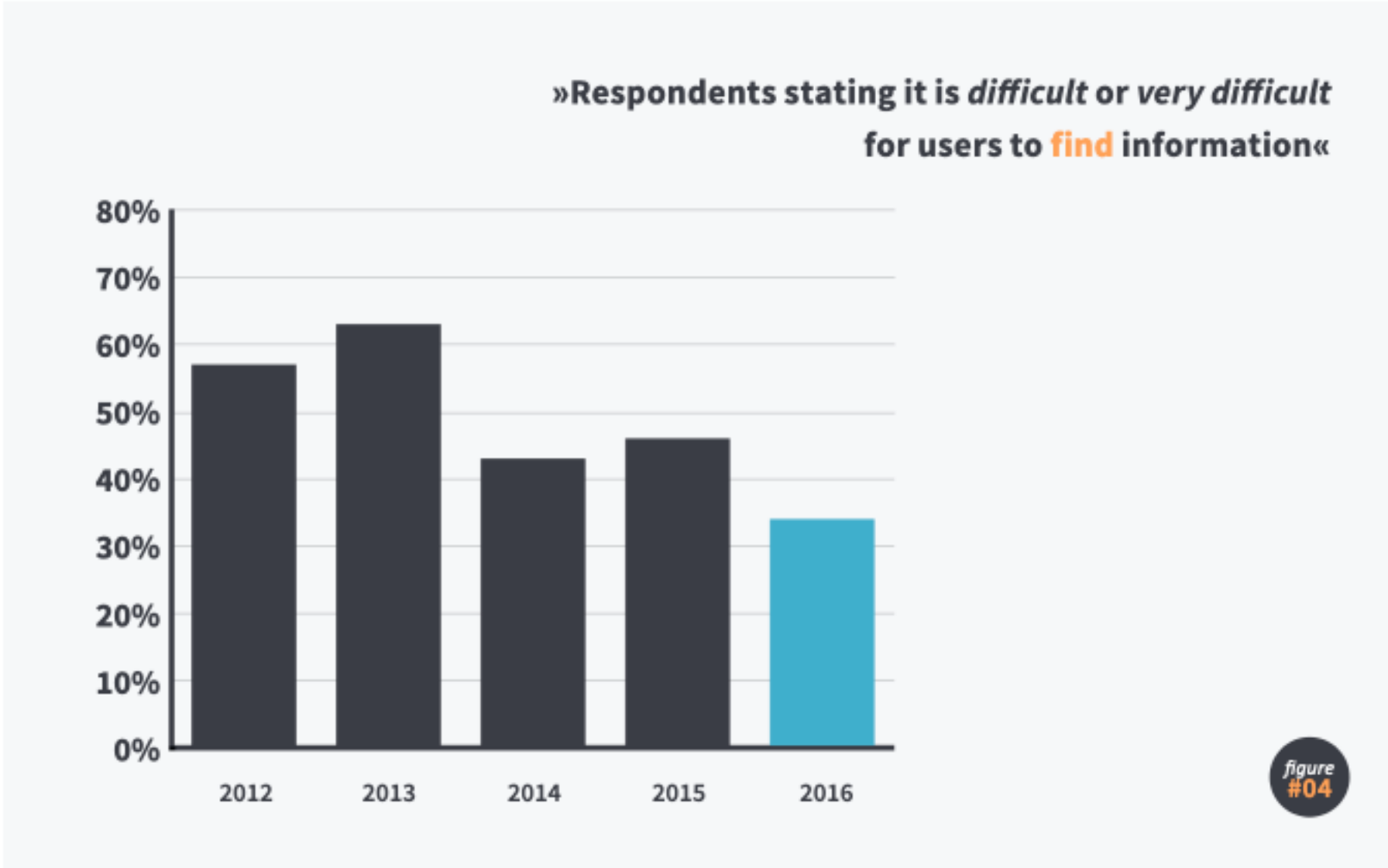
- Cross organisational responsible
- Information quality responsible
- Search analytics

TECHNOLOGY

- Aggregated content sources
- Search-as-a-service architecture
- Targeted search applications

figure #03

https://findwise.com/en/Enterprise-Search-Findability-Report-2016



https://findwise.com/en/Enterprise-Search-Findability-Report-2016

»Organisations with the ability to search across multiple content repositories (systems, databases) in one or more search application(s)«

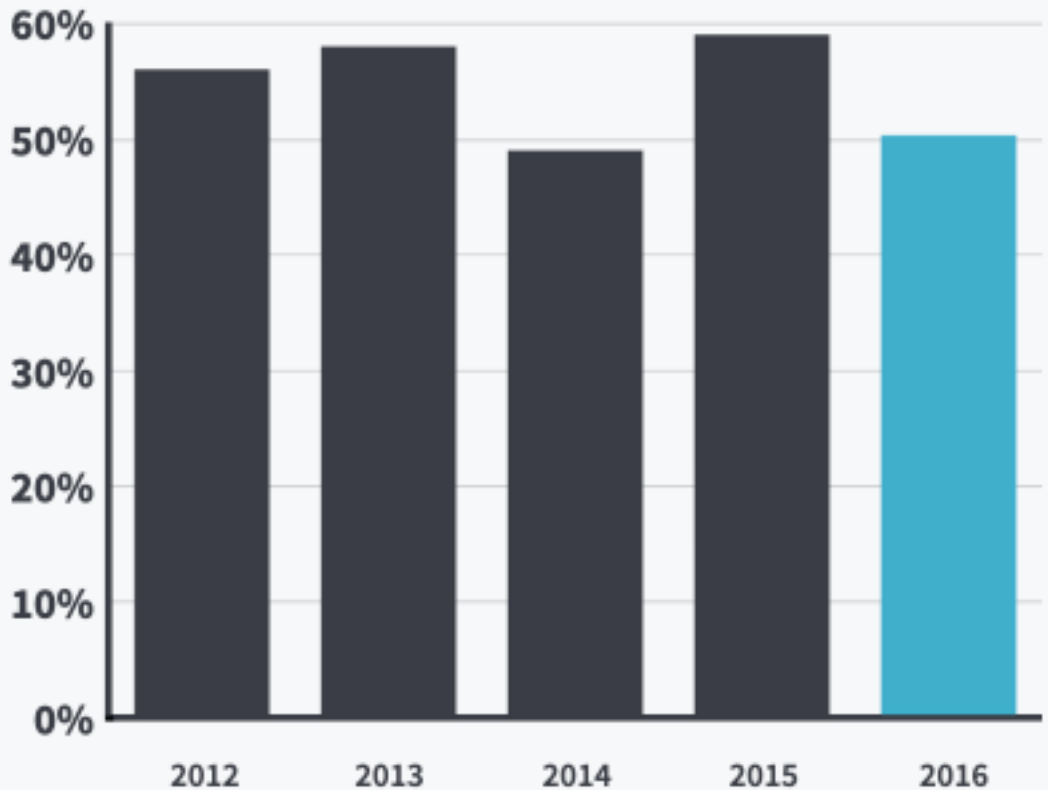
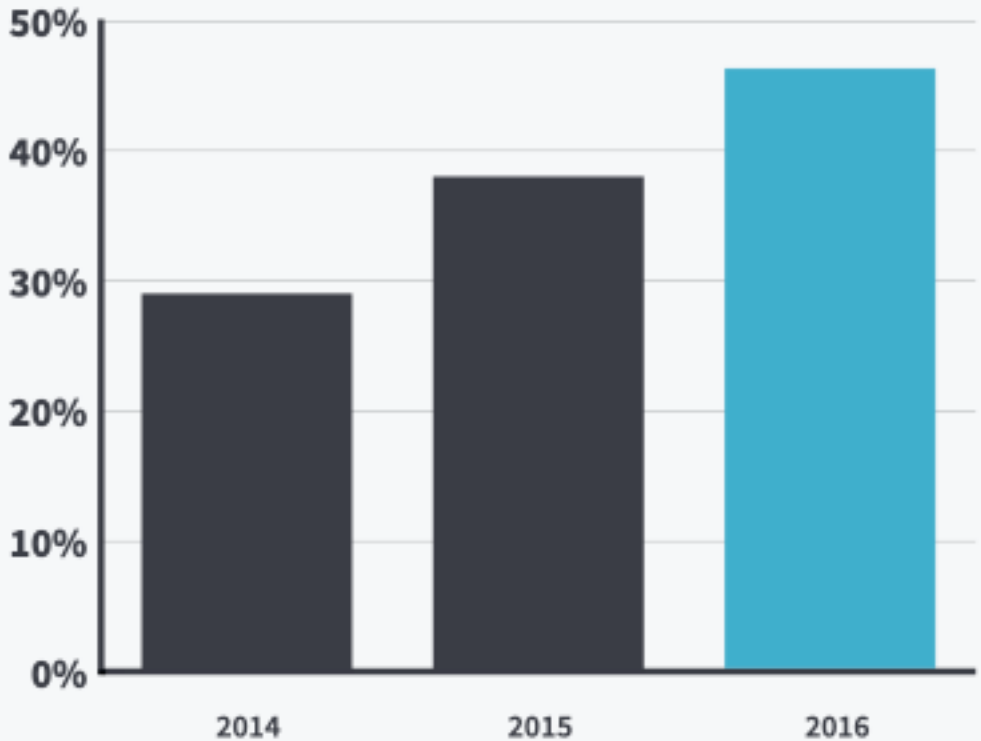


figure #06

https://findwise.com/en/Enterprise-Search-Findability-Report-2016



»Organisations planning to replace existing search technologies within 3 years«

figure #08

https://findwise.com/en/Enterprise-Search-Findability-Report-2016

»Organisations providing user guidelines on what metadata to add when publishing content«

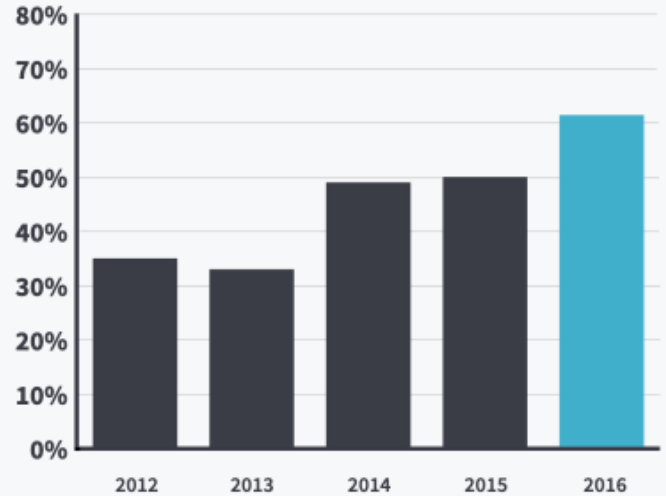
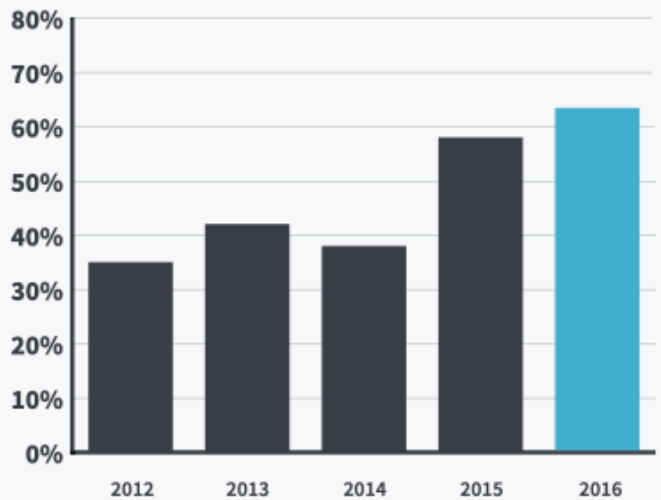


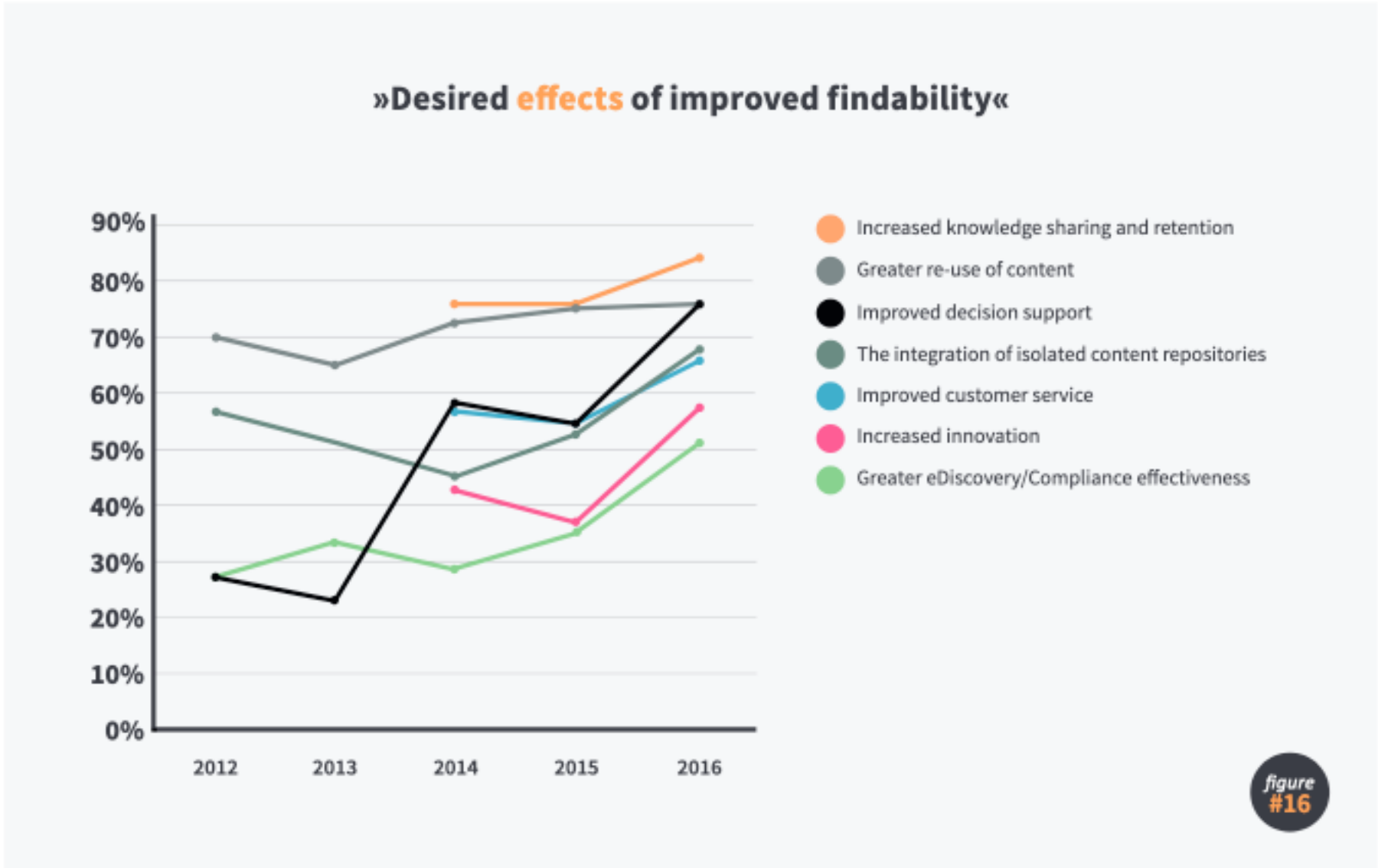
figure #09



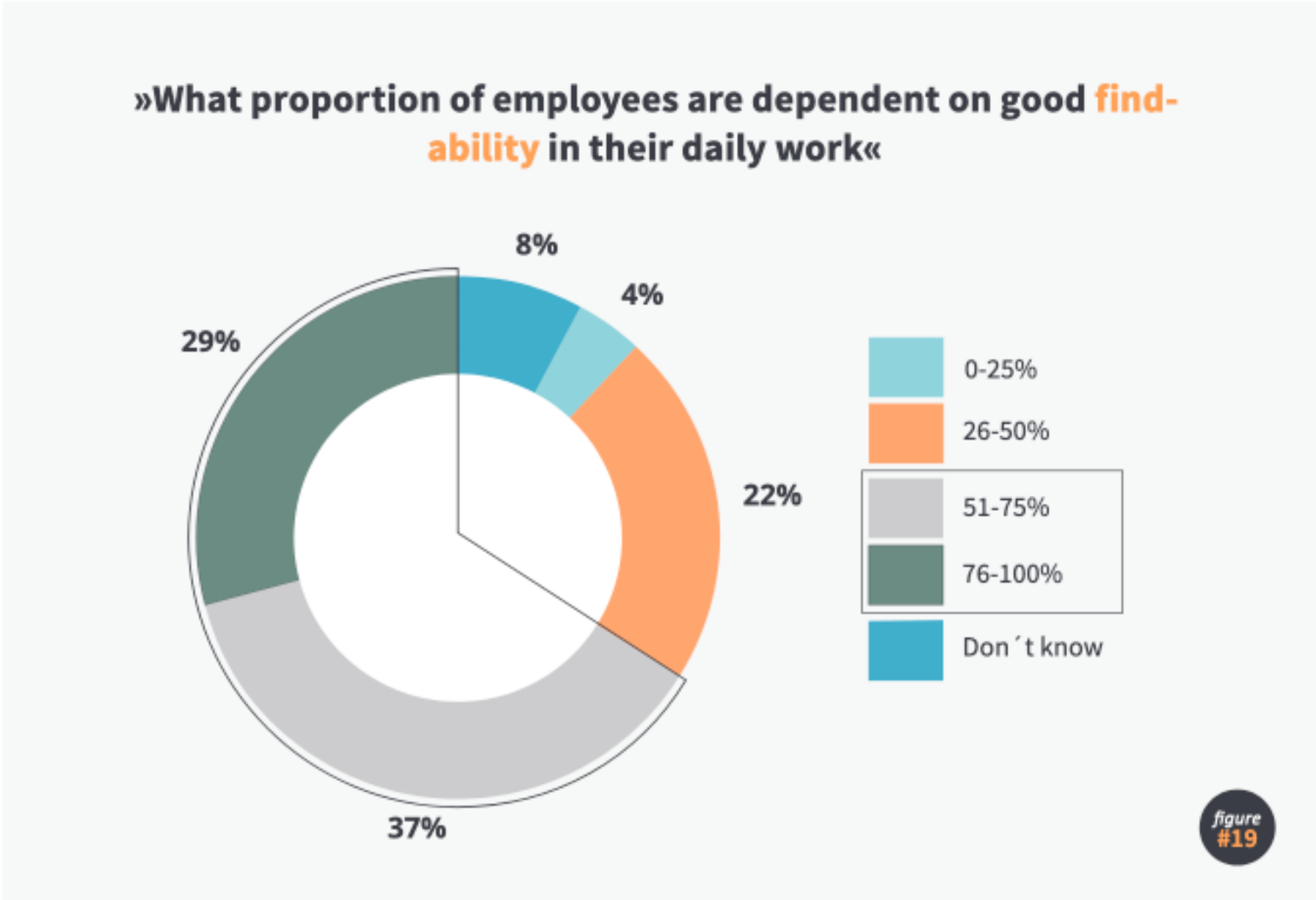
»Organisations using taxonomies to describe, organise and classify content«

figure #10

https://findwise.com/en/Enterprise-Search-Findability-Report-2016



https://findwise.com/en/Enterprise-Search-Findability-Report-2016



https://findwise.com/en/Enterprise-Search-Findability-Report-2016

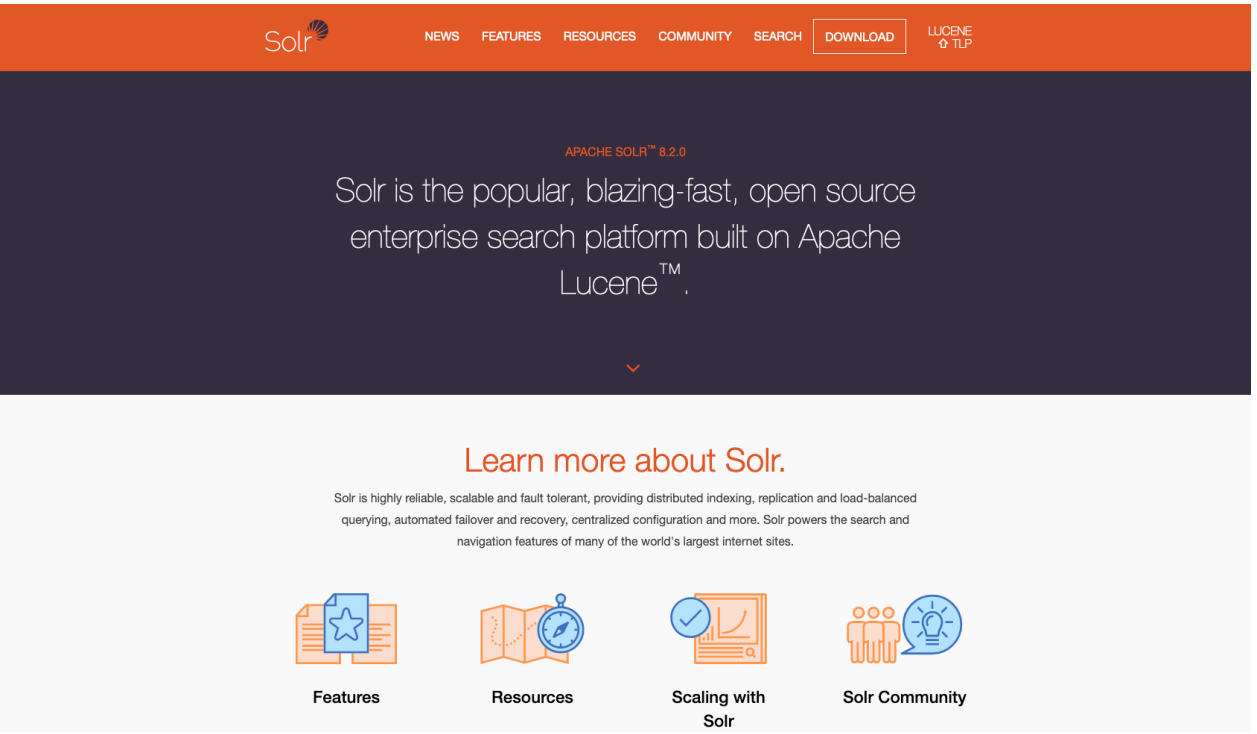
Figure 30 illustrates the information areas respondents specify as most important. People is the second most wanted information and finding contact details is important in most organisations.

https://findwise.com/en/Enterprise-Search-Findability-Report-2016

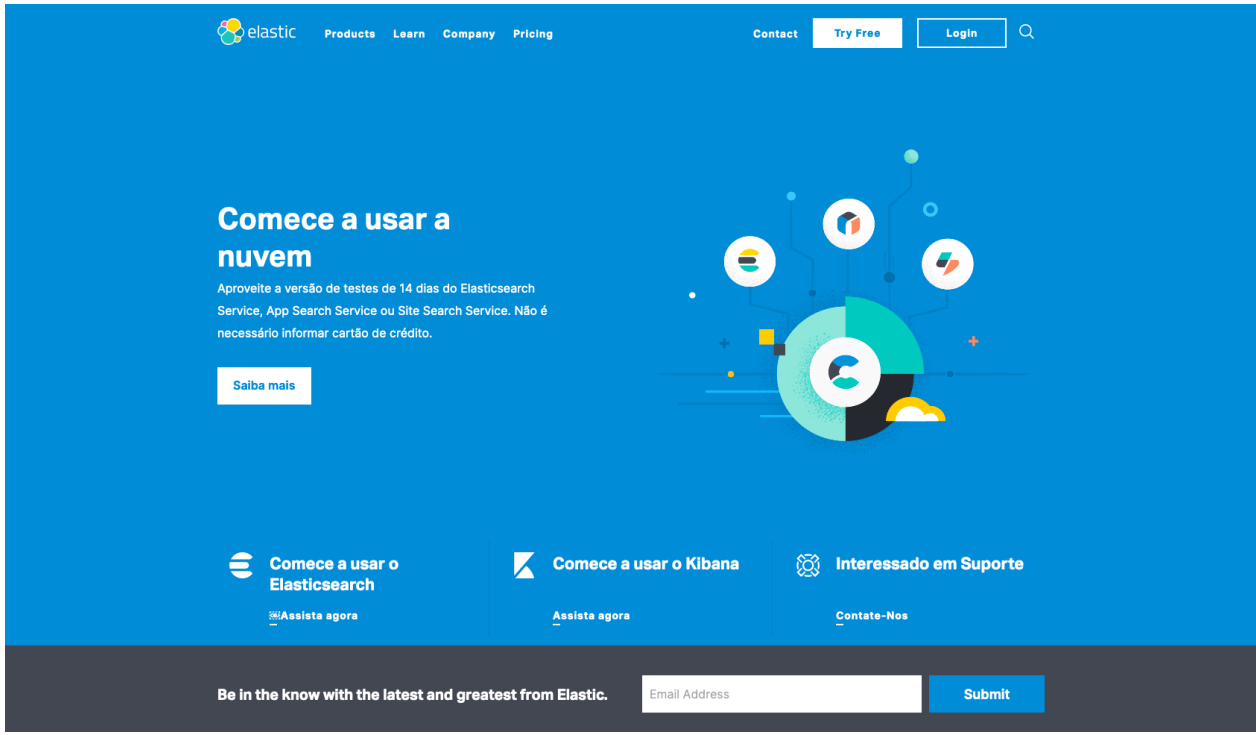
Algumas referências de recomendação

- http://intranetfocus.com/wp-content/uploads/2019/03/Search-Insights-2019-from-The-Search-Network.pdf
- http://www.flax.co.uk/wp-content/uploads/2018/03/Search-Insights-2018-from-The-Search-Network.pdf
- https://www.manning.com/books/relevant-search

Oportunidade de pesquisa



https://lucene.apache.org/solr/



https://www.elastic.co/pt/