

Go with the dataflow

Feasibility of using Internet as a data source: methods, case studies and indicators

Pim den Hertog (Dialogic), Reg Brennenraedts (Dialogic), Robbin te Velde (Dialogic), Christiaan Holland (Dialogic), Ronald Batenburg (Utrecht University & Dialogic), Slinger Jansen (Utrecht University), Sjaak Brinkkemper (Utrecht University)



Background

Starting point of this R&D project is the idea that researchers, statisticians and policy-makers could benefit more from the wealth of data that can be derived directly from the internet. The project was initiated by Dutch Ministry of Economic Affairs and conducted by the research consultancy Dialogic and the Utrecht University, both situated in The Netherlands.

The research had two goals:

1. Identify new data and indicators derived directly from the internet to map and describe new phenomena associated with the Emerging Digital Economy that cannot be covered using regular statistics
2. Explore and assess the usefulness of less invasive methodologies for deriving new and substitute data and indicators for the Emerging Digital Economy

Overall we assessed the value added of using these innovative methods for describing the emerging digital economy.

The concept of digital footprints are essential in this research. If people use the internet, they almost always leaves traces of their activities. Analogue to the environmental footprint, these traces are called digital footprints.

The Emerging Digital Economy is the ongoing process of digitalization affecting economic processes in firms, markets and industries and at the macro-economic level. It has an impact on the both the new and old economy, and both final and intermediate markets.

	INFORMATION	ORDERING	PAYMENT	FULFILLMENT & LOGISTICS
B2B	B2B online music marketplace	The 'Big Four' (Sony BMG, EMI, Universal, Warner) Wholesaler of music on CD's and MP3's		
		Physical music stores, e.g. Free Recordshop, Music Store, Media Markt		
B2C		Online music stores, e.g. Bol, Proxix		
		Online sale of ring tones, e.g. Jamba, Boltblue		
C2C			Provider of online financial services, e.g. Paypal, Ideal	Charts, e.g. Top-40, Top-50 Postal services (UPS, TNT)
	Online marketplaces, e.g. eBay, marktplaats, speurders			
	Internet search engines, e.g. Google, Yahoo			Torrent trackers e.g. TorrentSpy, TPB
	Sociale networks, fan sites, LastFM	Online storage, e.g. MegaUpload, RapidShare		Online storage e.g. MegaUpload, RapidShare

Focus points

To obtain insight in a certain phenomenon or a sector, one has to know where the interesting data can be gathered. We constructed a framework, which can be seen in the picture above, that allowed us to analyze a sector in a structural manner. By using a value chain approach and a market differentiation twelve cells arise. The blue rectangles show the focus points. To answer a question regarding properties of a sector or phenomenon, one has to take a look at the framework and analyze which focus points could be useful. The example above shows us were data regarding the music sector can be found. For example, suppose we want to know the most popular song at this moment. We could take a look at the website of an organization that publishes charts. But we could also create a direct link with the ERP systems of the largest record companies or wholesalers of CD's. Moreover, maybe we could also find this data by looking at the online transactions made. And in the illegal segment, we could take a look at Torrent Trackers or one-click storage websites.

The case studies

We conducted eight case studies: Pigs, housing, webstores, recorded music, product software, internet TV, gaming and social networks. The case have much variation regarding the level of digitalisation and are both in the 'old' and 'new' economy.

In these case studies we applied the framework displayed above to find focus points for data. Next, we gathered existing data that used internet as a source of data. For a number of cases we also conducted real experiments. The outcome of each case study in a figure analogue to the housing case study displayed on the right.

Next to the research we already did, we have some suggestions

User centric

Panel for use of marketplaces
Monitor use of internet-TV

Network centric

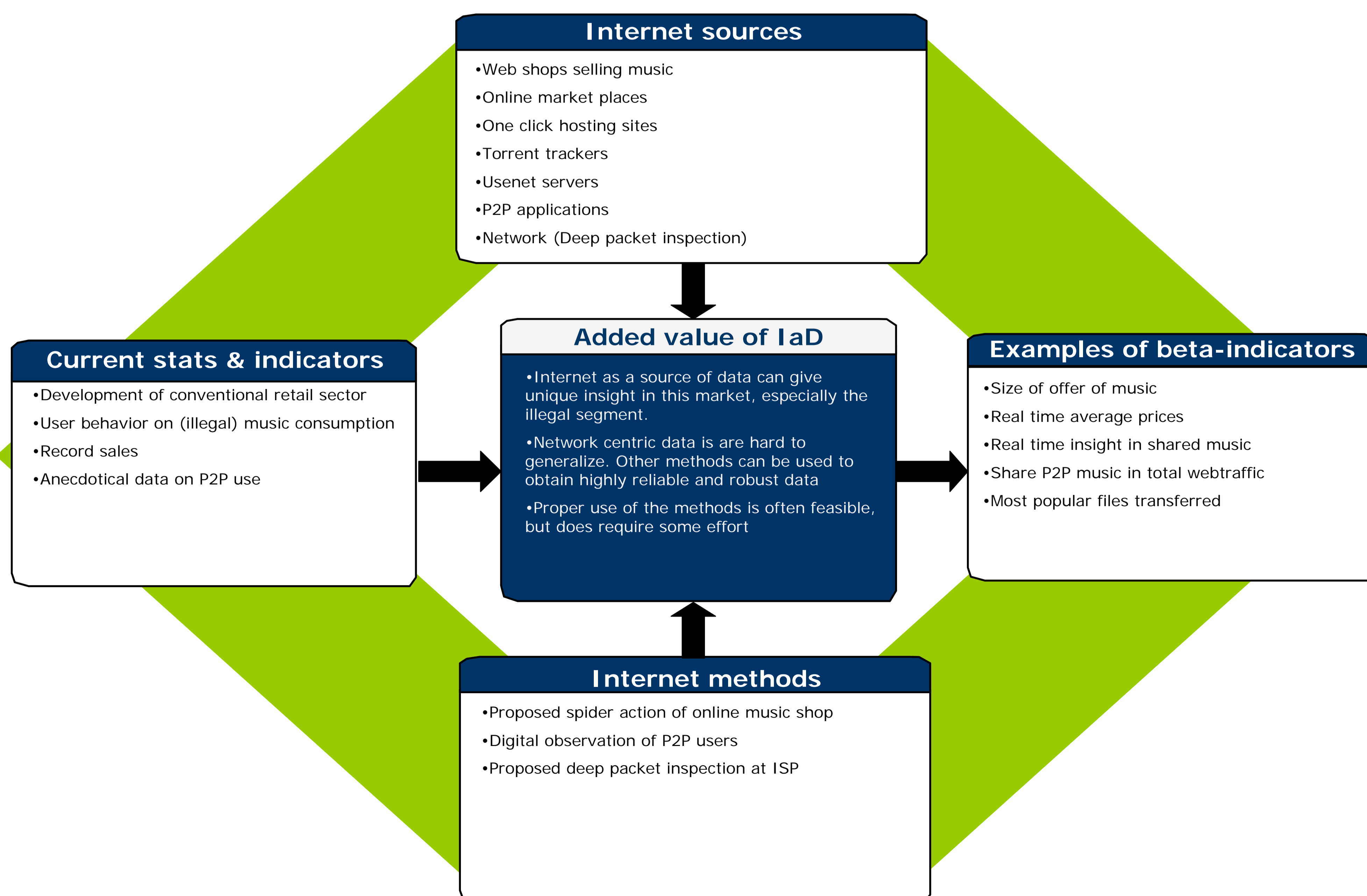
Measure use Citrix
Measure use DRM

Measure use instant messaging

Site centric

Real time housing prices
Occurrence long tail at web shops

Spider SNS on demographic properties



Overall conclusions

- IaD-methods are helpful in mapping new trends, phenomena and markets that can not be tracked with existing methods and statistics
- The potential of IaD-methods for substitution of existing indicators and statistics should not be overestimated
- IaD methods may lead to a new category of beta-indicators for the emerging digital economy
- Some typical trends in the emerging digital economy were confirmed in the case studies
- There is no "one size fits all approach" for measuring the emerging digital economy
- The value of Internet as a data source is highest in market with particular characteristics
- The practical and statistical usability of IaD is the result of various trade-offs

Considerations:

- Faustian bargain: trade-off between efficiency, objectivity, timeliness and cost-effectiveness on the one hand and validity and privacy on the other hand
- Sometimes there simply are no alternatives to the use of IaD methods

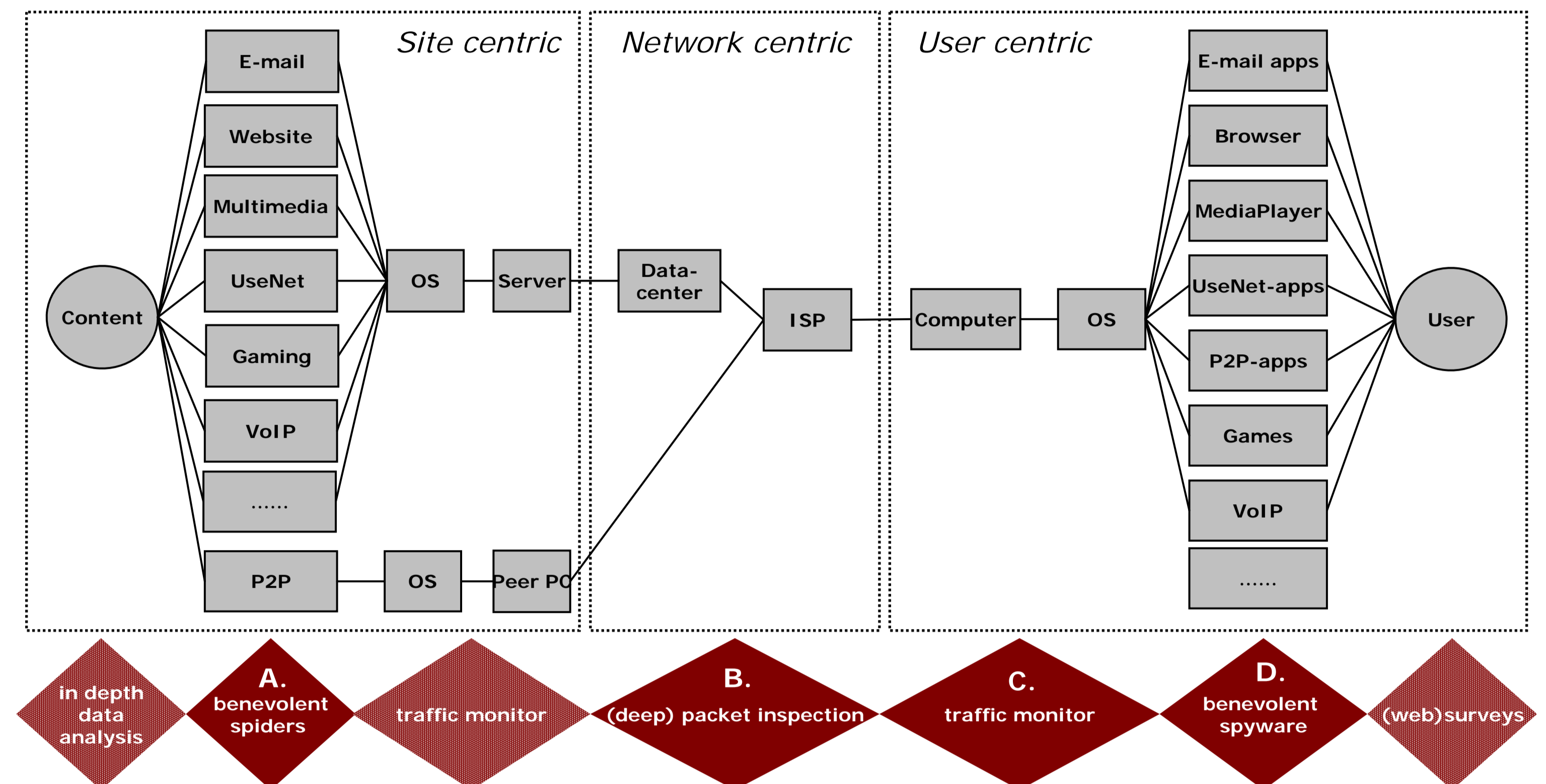
How we see these beta-indicators:

- There is a new category of beta-statistics that are especially suitable to pick up early trends in the EDE.
- We should assess the practical and statistical quality of these.
- If these statistics do not meet the quality standards of beta-statistics they should be dropped.
- The quality of the remaining beta-statistics should be improved
- Some beta-indicators promoted to the league of regular statistics.

Finding data on the Emerging Digital Economy by using the internet

The figure above shows the way users (right side) are able to obtain content (left side). Digital footprints are left on every step of the way. There are three fundamentally different measurements:

1. *Site centric measurements* are conducted in the vicinity of the content
 - a. *Benevolent spider*, measuring the properties of a single set of content
 - b. *Deep packet inspection*, measuring internet traffic between users content (all users, all applications)
2. *Network centric measurement* are conducted on the network (internet) connecting the user and content
 - a. *Traffic monitor*, measuring how a single person uses his computer (one user, all applications)
3. *User centric measurement* are conducted in the vicinity of the user
 - a. *Benevolent spyware*, measuring how a single person uses an application (one user, one application)



Lessons learned: Content

- Information on goods & services equals economic value in itself
- Digital footprints are substantial and likely to increase
- Market characteristics are key for usability IaD
- Quite simple questions can be enlightening and might result in beta-indicators
- Digitalization contributes to more fuzzy barriers between markets, actors and between social and economic realm
- Peer-production & increased transparency most visible economic impact digitalization

Lessons learned: Method

- Absence of an one size fits all approach
- Usability depends on: (a) Digital information on products and services, (b) Digitalization of the products and services themselves
- Difference between technical and practical availability (privacy, need for co-operation third parties)
- Often no clear demarcation of sectors
- Added value IaD mostly in signaling new trends, developments and phenomena
- Peer-production & increased transparency most visible economic impact digitalization

Lessons learned: statistical usability

- Efficiency: IaD is 'at the right place at the right moment'
- Objectivity: IaD does not (directly) use a human interviewer
- IaD-data is based on spontaneous behavior
- Reliability: Automated (IaD) measurements repeatedly return the same results
- Validity: Non-panel based IaD methods very hard not covered true existing methods and statistics
- Validity versus efficiency and objectivity

	Robustness (internal validity)	Representativity (external validity)	Transparency	Longitudinal use
Spyware and Traffic monitor	High	High	Very High	High
DPI at ISP	High	Low	Very low	Medium
Benevolent spiders	Low-medium	Varies	Medium	Low