Differences in Data-Sharing Attitudes and Behaviours

Flavio Bonifacio[1]

Abstract

This article reports the results of a survey conducted between 18th November and 18th December 2017 about different aspects of data sharing. After a short description of the data gathering task, the report describes the sample, the univariate distribution of the most important variables related to the work of data archiving and the attitudes concerning the data sharing activity: problems encountered, propensity to share the data, satisfaction obtained. Part of the report illustrates models suitable for interpreting the results and finally gives some advice for promoting data services. Some international comparisons of the results are proposed in the annex.

Background

In October 2017 Tuomas Alatera began a mail discussion among members of IASSIST about data sharing, writing the email '*I would share the data but…'*. Following this starting point many suggestions came from IASSIST members (see annex 2).
In November 2017 at the University of Turin a seminar was held on *Data Archiving*, *Dissemination and Reuse. A backward sight to go ahead* which addressed many similar questions.
This CAWI survey was conducted between 18th November and 18th December 2017 and reached 83 people working in the field of data curation and data analysis across the world (the sampling list used almost 500 email addresses of IASSIST members, 69 email addresses of EDDI17[2] (European Data Definition Initiative, 17th Congress, Lausanne, 5-6 December 2017) participants and almost 500 participants at the Turin seminar and from our mailing list. The questionnaire was built using selected questions from the email exchange among the Members of IASSIST cited above and already used in the survey *Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide*[3].
The questions regard different aspects of data sharing: tools used in building metadata, problems encountered in order to share the data, the propensity to share the data, and the satisfaction obtained from different working tasks.
The problem we tackled was at first no more than a simple curiosity, *'Do there exist any differences in the attitudes about Data Sharing between Italians and not Italians'*? which soon became a more exciting questions file. Another idea came into my mind. Given the decreasing amount of the traditional financial resources perhaps it is now necessary to sell these services to a wider and renewed market.

 Bonifacio, Flavio (2018) Differences in Data-Sharing Attitudes and Behaviours, IASSIST Quarterly 42 (3), pp. 1-40. DOI: https://doi.org/10.29173/iq912

Survey objectives

The goal of this work is to analyse attitudes toward data sharing in order to find strategies to expand data sharing and to show the best paths to gain new 'Premium Followers', as we call the best performers in Data Sharing (below). This pragmatic point of view comes from empirical evidence that is compliant with other more detailed analysis coming from wider surveys. We confirm what other sources say: 'Results show that researchers in different regions have different perceptions about data and different data behaviours'[4]. Although in literature there are many other well documented perspectives from which to examine Data Sharing, covering a range of aspects[5] from the political to the technical, we think that one of the most pressing issues is how to expand best practices in sharing data.

There are many purported motives getting in the way of data sharing: from modesty to ownership, from sabotage to fear.[6] Every sort of excuse is used to justify the dislike of data sharing. An overview of several collected opinions[7] would be needed to get a more precise idea of the topic.

This paper will focus on how to target the market to sell services related to data sharing:

1. *Is there a target for selling data-sharing services among data users?*
2. *Are there any personal perspectives and attitudes that may influence data services diffusion?*

In order to answer these questions, we will:

1. Examine the one-way variable distributions for background variables to describe the sample
2. Examine the one-way variable distributions for attitudes variables
3. Combine attitudes variables via PCA analysis to build indexes (work satisfaction index, sharing problems index, sharing propensity index)
4. Build a typology from the indexes
5. Analyse bivariate models with Country of origin as independent variables and attitudes variables, indexes and typology as dependent
6. Analyse bivariate models with other background variables (gender, age, study title and work sector) as independent variables and indexes as dependent
7. Analyse a three variables path with Country of origin as exogenous variable, Working sector as endogenous variable and propensity index as dependent variable
8. Analyse bivariate models with background variables as independent variables and metadata use as dependent to reinforce the hypothesis of existence of differences among data users
9. Definition of target variable for services promotion activity
10. Build a logistic model to set the best predictors for the target.

Warnings:

1. The population of this survey are Data Users. The IASSIST member list, the list of participants at the meetings quoted above, are just the sources where the email addresses have been found, the sample lists. To be more precise our population is restricted to Data Users comprised in those lists.
2. In this paper we refer to a Country of origin variable, Italians and Not Italians. I did this not because I think Italians are so important in the modern Data World, but because I supposed they are less involved in Data Sharing practices and might be representative of other countries in the same condition. As the following report will show they may have different attitudes on Data Sharing. There are other reasons: there were also no funds and resources to expand the research, for example.
3. Look at the annexes: there are research operations that may be clarified only with a methodological in-depth analysis. Annex three about factorial analysis explains why I choose those factors I used.

4. Look at the dimension of the sample of this survey: we had 83 respondents. The confidence interval may become very large. Every effort to build theories at this point would be premature and presumptuous and not sufficiently data based. Instead, the results should work as a suggestion to continue the survey in a quantitative or qualitative manner.

The survey results

One-way distributions: background variables

First, who are the respondents to the questionnaire? Almost half of them work at a university, a quarter in private agencies, and one fifth in public administration or in a non-profit area (tab. 1).

Tab. 1 – Work Sector distribution

	N	*%*
University	40	48.2
Public administration	10	12.0
Private company	22	26.5
No profit	8	9.6
Other	3	3.6
TOTALE	83	100.0

Although the sample drawn from the sample lists[8] is not random, it is representative of data workers: all those interviewed are involved in the field of data analysis (at least as a user) or in data curation. The results must be viewed keeping this mind: we are dealing with a population of data field experts.
Among respondents there are slightly more men than women (tab. 2) and the most represented age cohort is 51-65 (tab. 3) who are the oldest ones. *Does this mean that the interest in Data Sharing is decreasing?*

Tab. 2 – Distribution by Gender

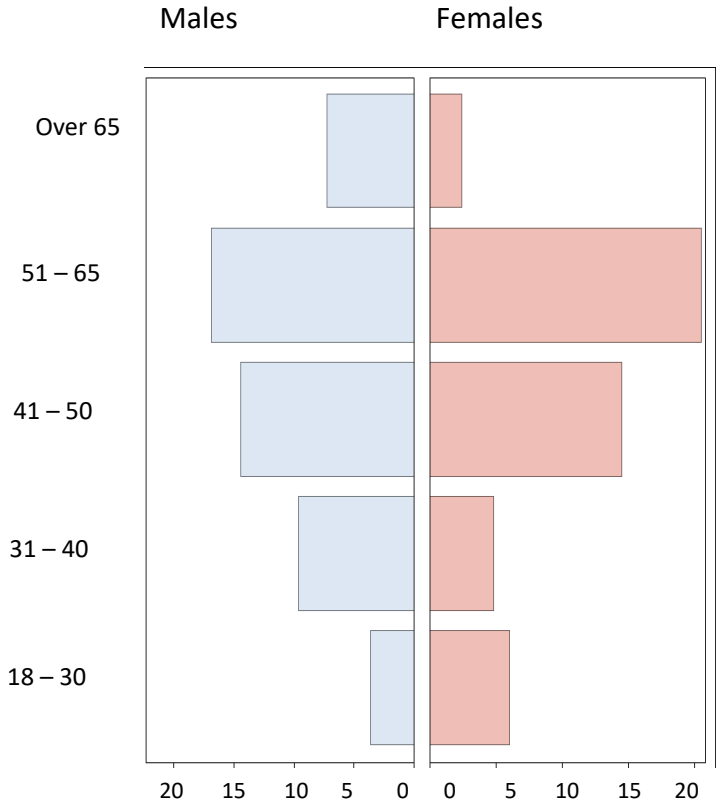
	N	*%*
Male	43	51.8
Female	40	48.2
TOTAL	83	100.0

Tab. 3 - Age distribution

	N	*%*
18-30	8	9.6
31-40	12	14.5
41-50	24	28.9
51-65	31	37.3
Over 65	8	9.6
TOTAL	83	100.0

The age pyramid of respondents is reported in fig. 1. Although there is not statistical significance between genders and age we note that females are more represented in the 18-30 and 51-65 age class and slightly more also in the 41-50 class.

fig. 1 – Age pyramid



As we can expect, the majority have a post-graduate qualification (*Master* or *Phd*) and almost all are at least graduates (tab. 4).

Tab. 4 – Distribution by study title

	N	*%*
High school	2	2.4
Graduation	34	41.0
Master	31	37.3
Phd	16	19.3
TOTAL	83	100.0

Half of respondents come from social studies and one third from scientific studies (tab. 5). As expected social studies are preeminent.

Tab. 5 –Study sector

	N	*%*
Humanities	13	15.7
Scientific	28	33.7
Social (sociology, psychology, economics)	42	50.6
TOTAL	83	100.00

One-way distributions: data sharing attitudes

As reported above, most of the questions have been extracted from *IMDSM2017* (see annex 2) and from the quoted survey *Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide* and grouped into three main conceptual frameworks:

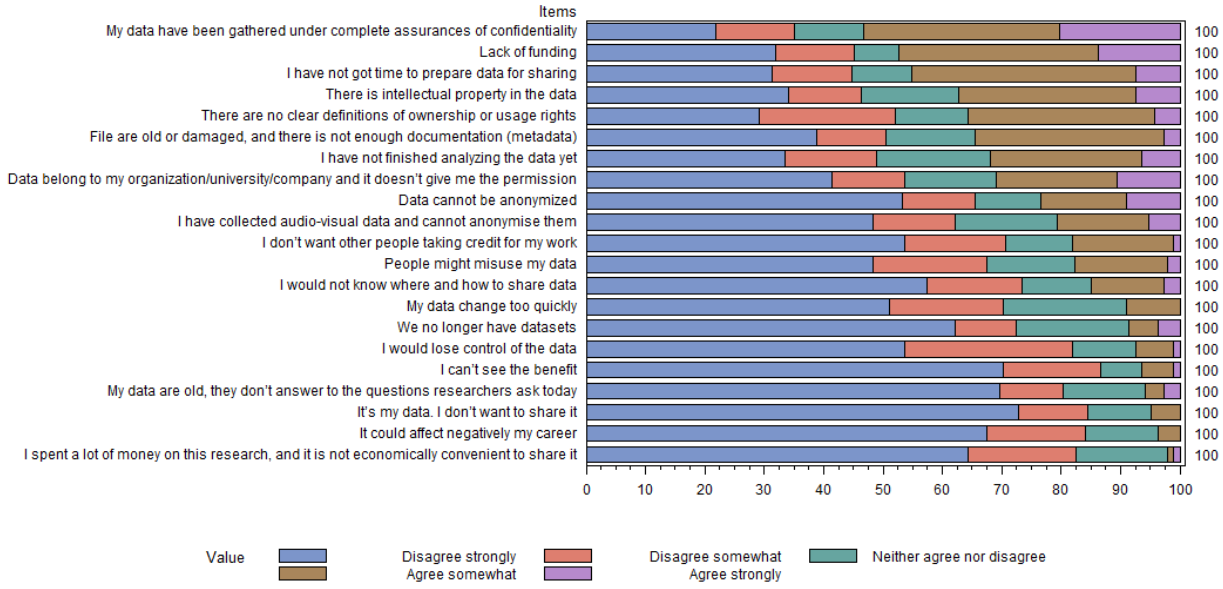
1. problems emerging in data sharing (21 items)
2. items influencing work satisfaction (7 items)
3. items influencing data sharing propensity (6 items)

To avoid bias due to oversampling of Italian experts, frequencies are weighted in such a way that the Italian subsample will lose weight in the figures. The solution is somewhat artificial and not optimal, due to the impossibility to calculate the effective weight of Italian experts over the entire data expert population. If we count as a proxy the number of Italian experts in the international organizations (i.e. Iassist) this number tends to be zero. Instead we have taken into account the conditional frequency of using metadata standards (the precise survey question is: What metadata standards do you currently use to describe your data?) given the subsample of not Italians as a weight. The underlying hypothesis is that those who do not use metadata standards are less involved in the data sharing activity. The weight is given by the fraction .74286/.42169 for not Italians and .25714/.57831 for Italians where numerators are the proportions of not Italians using (.74286) (or not using (.25714)) metadata standards and denominators the proportions of the two subsamples. As a result, Italians will weigh almost half and not Italians almost 1.76 times the real subsample weight. We will use this weight to present one-way frequencies of the attitudes shown by the people interviewed. This weight has no effect when we consider the conditional distributions given by the subsample type (not Italian/Italian subsamples).

Problems emerging in data sharing

Among the problems mentioned, those recognized as creating more difficulties are, in order of frequency: *confidentiality, lack of funding, lack of time, intellectual property, no clear definitions of ownership, my data are old and not sufficiently documented, I have not finished analysing the data yet, the data belong to my organization* (agree strongly, agree somewhat over 30%, confidentiality over 50%, fig. 2) (the benchmark value is 21.78, the marginal distribution value over all items of the considered values). Also mentioned are problems related to privacy and propriety, to resources (time and money), to data quality and documentation, and to data analysis.

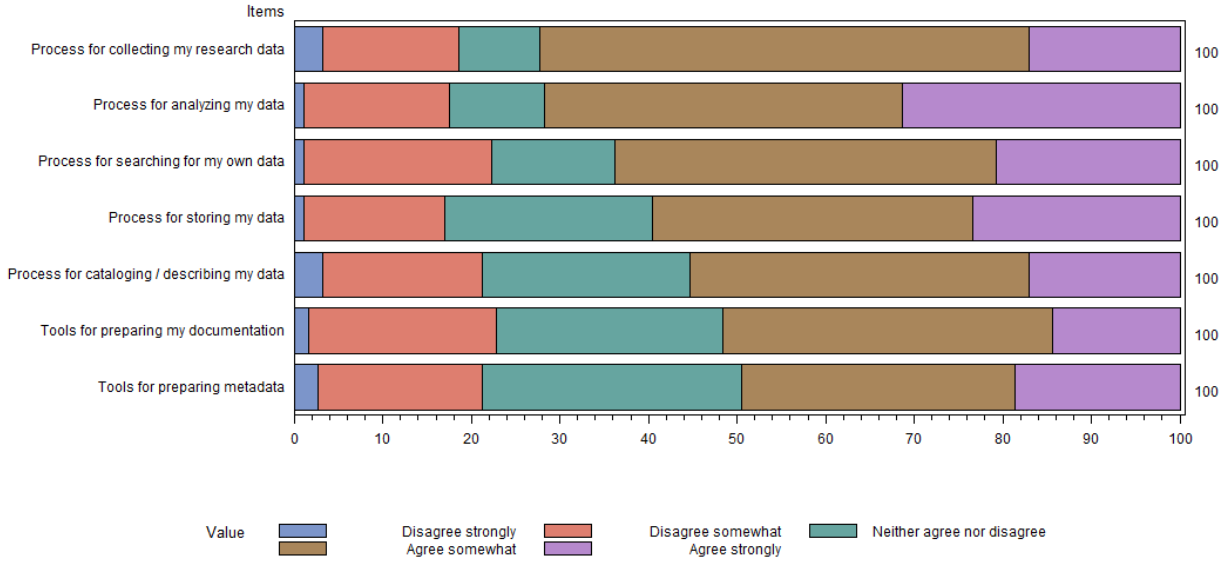
Fig. 2 - How much do you agree with these statements about the reasons that prevent data sharing? I WOULD SHARE THE DATA BUT…



Items influencing work satisfaction

Data gathering, data analysing and data searching are the items that most influence work satisfaction (all over 70%). *Documentation* (51%) and *metadata preparation* (49%, less than half) tools are the least satisfying (fig. 3, the benchmark value is 60.55, the marginal distribution value over all items).

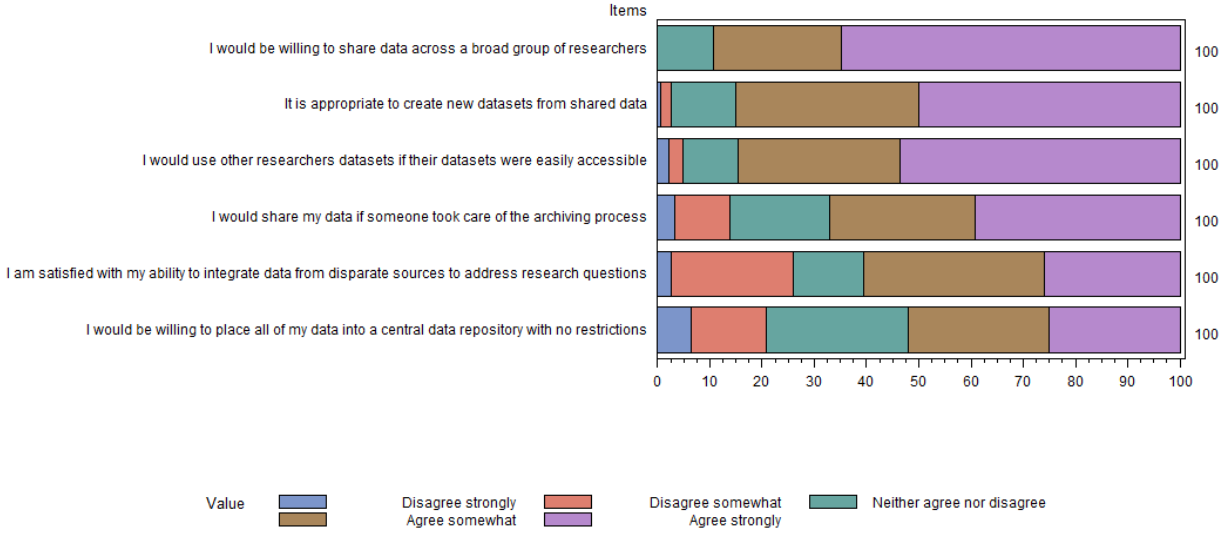
Fig. 3 - The following statements relate to how you collect and use research data. Tell us how much you agree with the following ways to complete this sentence: I AM SATISFIED WITH THE…



Items influencing data sharing propensity

The most appealing items stimulating the propensity to share data seem to be *sharing data among broad groups of researchers, creating new datasets from shared data and the ease of data access* (agree strongly, agree somewhat over 80%, fig. 4). The least cited item is *I would be willing to place all of my data into a central data repository with no restrictions* which equals almost half of the expressed preferences (52%) (the benchmark value is 73.11, the marginal distribution value over all items). This order is interesting because it leaves "My Data Sharing", the active part of Data Sharing, in the last position.

Fig. 4 - The following statements relate to sharing scientific data. Tell us how much you agree with each statement.



Three attitude scales

The items shown in the previous tables work very well together showing a very high reliability coefficient. We built three attitudes scales or indexes, one for every item battery: *WSI- Work Satisfaction Index* with a standardized coefficient (Cronbach's alpha) of *0.90, SPI-Sharing Problems Importance Index* with a standardized coefficient (Cronbach's alpha) of *0.91, SPN-Sharing PropeNsity Index* with a standardized coefficient (Cronbach's alpha) of *0.80*)[9].
The weighted distributions[10] show that satisfaction and propensity indexes are more concentrated over high scores: a high WSI score means that respondents find the tasks of their work more satisfying than the other respondents (39%, fig. 5), a high SPN score means that respondents are more active in data sharing (41% fig. 7), a high SPI score means that respondents think that there are problems in sharing data (24%, fig. 6). As we can see only SPI has the lowest frequency for high score.

Fig. 5 – Work satisfaction index

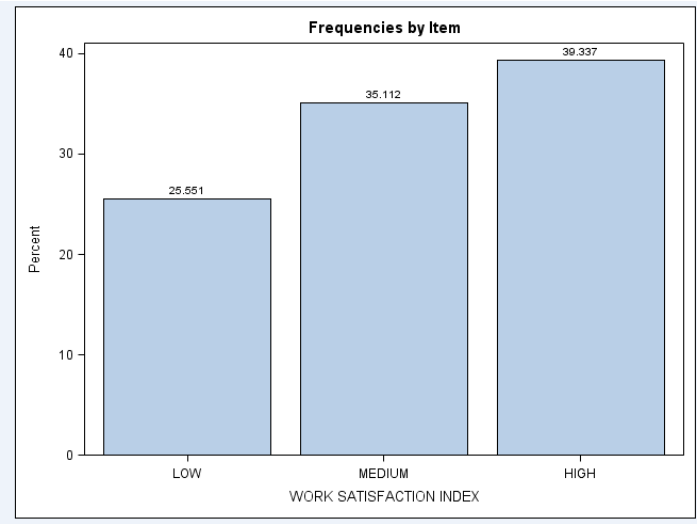


Fig. 6 – Sharing Problems Importance index

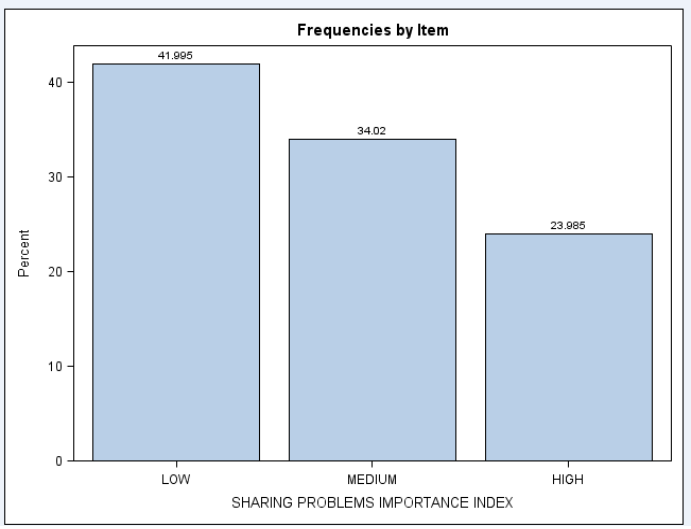
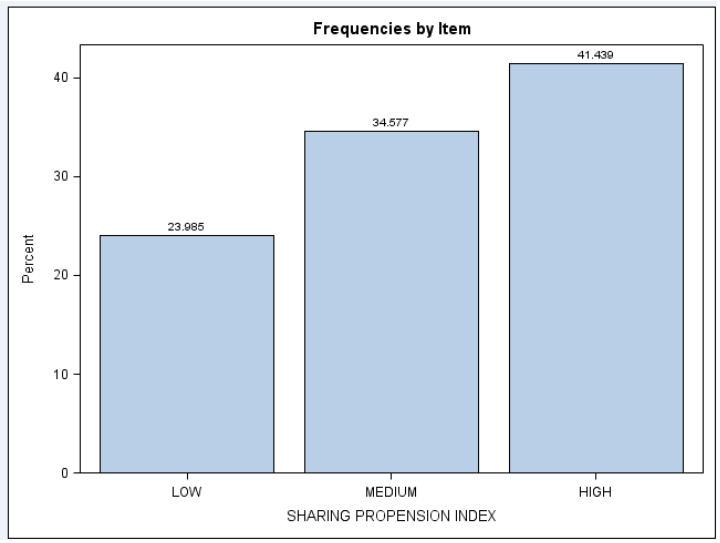
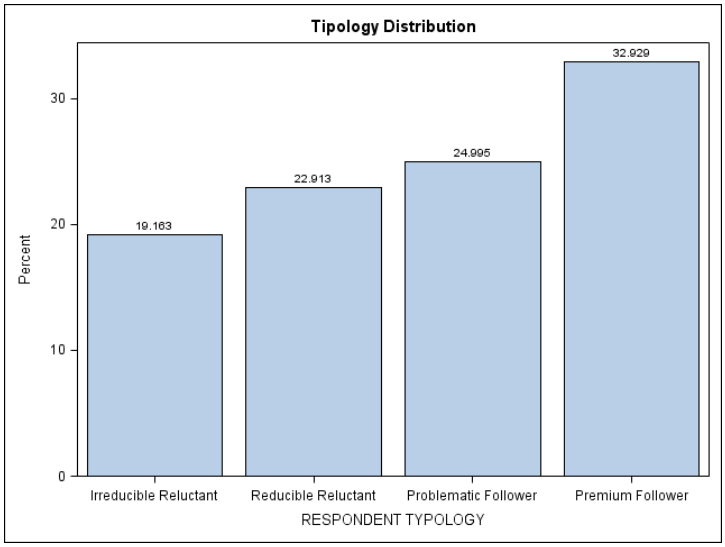


Fig. 7 – Sharing Propensity Index



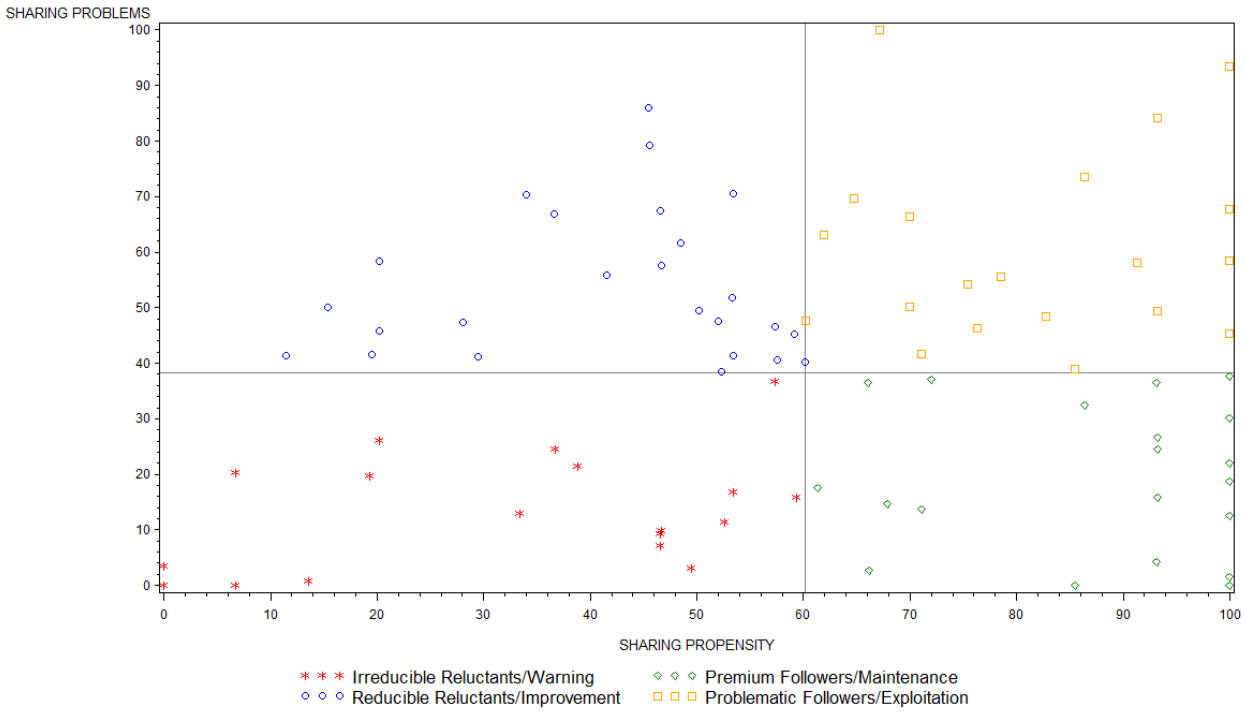
If we look at the correlation among the indexes we find that they are almost uncorrelated, although they come from different PCA analysis. That means that a high propensity to share data does not influence the attitude related to problems emerging in data sharing and *vice versa*, and neither of them influence the satisfaction index. The indexes measure truly different views of the respondents.
Moving the cut-off point of the index distributions to the mean[11] and therefore reducing the numbers of the index categories from three to two, and combining the resulting values, the following figure shows a well-balanced distribution (fig. 8). The first[12] category means low sharing propensity and few sharing problems (*Irreducible Reluctant)*, the second category means low sharing propensity and high sharing problems (*Reducible Reluctant)*, the third category high sharing propensity and many sharing problems (*Problematic Follower[13])*, the fourth category high sharing propensity and few sharing problems (*Premium Follower)*. *Premium Followers* are the most represented category (almost one third) in the weighted distribution.

Fig. 8 – Distribution by respondent's typology



What those categories exactly mean comes from the PCA analysis cited above and more precisely from the correlation matrix between factors and items used in building scales. Those matrices are reported here with some comments in appendix 3. Here just an explanation of the meaning given to the typology labels is provided: *Irreducible Reluctants* because they show low propensity in Data Sharing and do not recognise the related problems; *Reducible Reluctants* because they too show low propensity but have a feeling of the problems (perhaps they have a low propensity because of this feeling); *Problematic Followers* because they have a high level of propensity but also a high perception of problems (it seems that something is missing for them); *Premium Followers* because they have high propensity and do not perceive so many problems (presumably because they have solved them). If we think at the quadrant of customer satisfaction surveys, usually built combining importance and satisfaction of the items surveyed, each labelled action may be adapted to our typology: *warning* for *Irreducible Reluctants*, *Improvement* for *Reducible Reluctants*, *Exploitation* for *Problematic Followers*, *Maintenance* for *Premium Followers*. In fig. 9 is shown our magic quadrant reporting the scores obtained by each respondent for both of the measures used by the typology. Each point counts as one (it is not weighted).

Fig. 9 – Data Sharing Magic Quadrant



Models

In this section some significant relations between the most important variables described above will be shown. The sample is small, so it is not possible to work on the model with more than one or two independent variables[14]. Nevertheless, we have some interesting suggestions.
First, there are the differences between the Italian (as an example of countries with low interest in data sharing, where data sharing is not so widespread) and not Italian experts interviewed. Then some interesting models considering the work place and other variables are presented.

First on the list, for each items battery, are the items where the P value of Fischer F (homogeneity of variance test) is less than .1 indicating that the relation between the item and the subsample type is significant. Tables 6, 7, 8 report the average of each item of the *sharing problems importance index*, coded from 1 (fewer problems recognized in data sharing) to 5 (more problems recognized in data sharing). We observe that for all the items where the statistic F is significant, and for the majority of the other items, *Not Italians* recognize fewer sharing problems (or assign less importance to the sharing problems) than *Italians*. The biggest differences (the ones listed first) are for the opinions on *data misuse,* on *economic convenience* and on *visibility of data sharing benefits* (all significant at the 1% level).

Tab. 6 – Items pertaining to the Problem Importance index ordered by ascending P value - Means difference between Not Italians and Italians

Items Label	*Total P Value*	*Not Italians*	*Italians*	*F statistic*	*p value F*
People might misuse my data	2.039	1.771	2.813	13.57	0.00041
I spent a lot of money on this research, and it is not economically convenient to share it	1.565	1.400	2.042	9.522	0.00278
I can't see the benefit	1.506	1.343	1.979	8.245	0.00521
I would lose control of the data	1.730	1.571	2.187	6.902	0.01029
My data change too quickly	1.879	1.714	2.354	6.414	0.01325
My data are old, they don't answer the questions researchers ask today	1.586	1.429	2.042	6.061	0.01594
There is intellectual property in the data	2.645	2.457	3.188	4.456	0.03786
My data have been gathered under complete assurances of confidentiality	3.166	2.971	3.729	4.427	0.03847
It could negatively affect my career	1.522	1.429	1.792	2.942	0.09012
I would not know where and how to share data	1.868	1.743	2.229	2.694	0.10459
It's my data. I don't want to share it	1.474	1.400	1.687	1.743	0.19051

Items Label	*Total P Value*	*Not Italians*	*Italians*	*F statistic*	*p value F*
I have not got time to prepare data for sharing	2.765	2.886	2.417	1.731	0.19197
Data belong to my organization/university/company and it doesn't give me the permission	2.464	2.343	2.812	1.641	0.20379
I don't want other people taking credit for my work	1.947	1.857	2.208	1.363	0.24640
There are no clear definitions of ownership or usage rights	2.586	2.514	2.792	0.702	0.40465
Lack of funding	2.840	2.914	2.625	0.574	0.45090
I have collected audio-visual data and cannot anonymise them	2.155	2.114	2.271	0.220	0.64012
We no longer have datasets	1.777	1.743	1.875	0.210	0.64819
I have not finished analyzing the data yet	2.559	2.543	2.604	0.032	0.85819
Files are old or damaged, and there is not enough documentation (metadata)	2.479	2.486	2.458	0.006	0.93662
Data cannot be anonymized	2.138	2.143	2.125	0.002	0.96063

Not significant differences exist on the *satisfaction index*, where the average values are very similar between the two groups. There are no significant differences at the 10% level. The nearest values are on the *cataloguing process*. Nevertheless, Not Italians show a little bit more satisfaction for every aspect.

Tab. 7 – Items pertaining to the Satisfaction Index ordered by ascending P Value - Means difference between Not Italians and Italians

Items Label	*Total*	*Not Italians*	*Italians*	*F statistic*	*p value F*
Process for searching for my own data	3.611	3.686	3.396	1.155	0.28563
Tools for preparing my documentation	3.414	3.486	3.208	1.148	0.28705
Process for analysing my data	3.845	3.914	3.646	0.980	0.32524
Tools for preparing metadata	3.441	3.486	3.313	0.406	0.52594
Process for collecting my research data	3.675	3.714	3.563	0.338	0.56262
Process for storing my data	3.649	3.686	3.542	0.299	0.58622
Process for cataloguing/describing my data	3.478	3.514	3.375	0.264	0.60908

For every item of the *sharing propensity item scale* the propensity is higher for *Not Italians*, especially for the *willingness to share data among a broad group of researchers* and for the *creation of new datasets moving from already shared data*, both below 1% of significance.

Tab*. 8 – Items pertaining to the Sharing Propensity Items ordered by ascending P Value – Means difference between Not Italians and Italians*

Items Label	*Total*	*Not Italians*	*Italians*	*F statistic*	*p value F*
I would be willing to share data across a broad group of researchers	4.541	4.743	3.958	27.63	0.00000
It is appropriate to create new datasets from shared data	4.318	4.457	3.917	7.553	0.00738
I would use other researchers' datasets if their datasets were easily accessible	4.313	4.429	3.979	3.886	0.05210
I would share my data if someone took care of the archiving process	3.893	4.000	3.583	2.144	0.14703
I would be willing to place all of my data into a central data repository with no restrictions	3.500	3.543	3.375	0.308	0.58038
I am satisfied with my ability to integrate data from disparate sources to address research questions	3.580	3.600	3.521	0.070	0.79240

The not surprising conclusion at this point is that the countries having an important Data Sharing tradition show fewer problems and a greater Data Sharing Propensity.

A synthetic overview

What is more exciting is to analyse the differences with a more synthetic view directly over the indexes, instead of analysing each item, as we have done above.
While the *Work Satisfaction* (0 not satisfied at all, 100 fully satisfied) *index* does not show significant differences by country, although *not Italians* seem to be more satisfied, significant differences exist for *Sharing Problems Importance* (0 Not important at all, 100 highly important) and Sharing Propensity (0 Not important at all, 100 highly important). *Not Italian*s recognize fewer problems with data than Italians and are more willing to share it (tab. 9). This fact is probably related to the lower data sharing propensity showed by Italians. We noted above that data sharing propensity and sharing problem importance are uncorrelated. That is true. But if we run a model with sharing problem importance as dependent variable, sharing propensity as independent and Country of Origin as controlling variable, the regression shows significant estimates for the Italian group with a positive correlation between sharing problem importance and sharing propensity. This relationship may be interpreted as an effect of problems encountered on the sharing propensity: the ones that have a higher sharing propensity recognize more problems. This may be counterintuitive, but may be plausible. Only those who work in the field may see the difficulties.

Tab. 9 – Means difference over the three indexes: Satisfaction, Problems Importance and Sharing Propensity by country of origin

	Work Satisfaction		*Sharing Problems Importance*	*Sharing Propensity*
		F=0.97 P(F)=0.3265	F=5.66 P(F)=0.0197	F=8.99 P(F)=0.0036
	N	*Mean*	*Mean*	*Mean*
Country				
Italians	48	60.84	43.98	51.84
Not Italians	35	65.98	30.62	71.70
Total	83	63.01	38.34	60.21

Following our classification, *not Italians* are more frequent among followers (premium and problematic), while Italians are more present among reluctant (irreducible and reducible) (tab. 10). The table is significant at an alpha level of 5%.

Tab. 10 – Means difference by typology and country of origin

	Typology				*Total*	
	Irreducible Reluctant	*Premium Follower*	*Problematic Follower*	*Reducible Reluctant*		
	%	*%*	*%*	*%*	*N*	*%*
Country						
Italians	25.00	12.50	22.92	39.58	48	100.0
Not Italians	17.14	40.00	25.71	17.14	35	100.0

Other Models

Given the dimension of the sample, it is hard to test a model with more than one or two independent variables. Looking at models with one independent variable, we meet four significant relationships. Gender is linked to the Sharing Problem Importance index at a significant level: women seem to recognize more problems than men (tab. 11). It would be interesting to study this relation in a more detailed way. Is it an indicator of a sort of digital (sharing) divide? Or is it an indicator of a "natural" more attentive female attitude? The data do not permit the verification of such a hypothesis. Further (qualitative?) studies would be needed on this topic.

Tab. 11 – Means by Index Type and Gender

	Work Satisfaction		*Sharing Problems Importance*	*Sharing Propensity*
	F=0.12 P(F)=0.7281		F=11.89 P(F)=0.0009	F=0.37 P(F)=0.5435
	N	*Mean*	*Mean*	*Mean*
Male	43	65.48	25.62	68.51
Female	40	63.89	41.97	64.79
Total	83	64.66	34.05	66.59

Although the older respondents seem to be more satisfied, to recognize fewer problems and share data more readily, age does not show a significant relationship with the measures tested (tab. 12). What is worth noting here is that Sharing Problems Importance is decreasing as the age increase, while the Sharing Propensity moves in the opposite direction: the Sharing Propensity increases with age.

Tab. 12 – Means by Index Type and age

	Work Satisfaction		*Sharing Problems Importance*	*Sharing Propensity*
	F=0.89 P(F)=0.4761		F=1.12 P(F)=0.3523	F=1.34 P(F)=0.2641
	N	*Mean*	*Mean*	*Mean*
18-30	8	62.49	43.14	58.16
31-40	12	65.21	27.61	61.95
41-50	24	59.24	31.95	60.71
51-65	31	66.49	38.33	71.52
Over 65	8	74.51	26.16	81.35
Total	83	64.66	34.05	66.59

The only significant effect of academic qualifications is on Sharing Propensity: the higher the qualifications, the higher the Sharing Propensity (tab. 13).

Tab. 13 – Means by Index Type and Study Title

	Work Satisfaction		*Sharing Problems Importance*	*Sharing Propensity*
	F=1.22 P(F)=0.3019		F=1.03 P(F)=0.3633	F=4.18 P(F)=0.0189
	N	*Mean*	*Mean*	*Mean*
Graduation	34	58.73	39.10	53.19
Master	31	66.71	30.62	67.94
Phd	16	67.84	35.69	78.72
Total	81	65.08	33.60	66.66

Working in an Academic or Public environment fosters Data Sharing Propensity too (tab. 14).

Tab. 14 – Means by Index Type and Work Sector

	Work Satisfaction		*Sharing Problems Importance*	*Sharing Propensity*
	F=0.26 P(F)=0.8570		F=1.84 P(F)=0.1470	F=8.93 P(F)<0.0001
	N	*Mean*	*Mean*	*Mean*
University	40	65.25	35.81	73.33
Public administration	10	59.83	27.52	77.76
Private company	22	61.32	42.34	45.33
No profit	8	64.58	21.81	39.97
Total	80	63.97	33.94	65.78

A More Complex Model

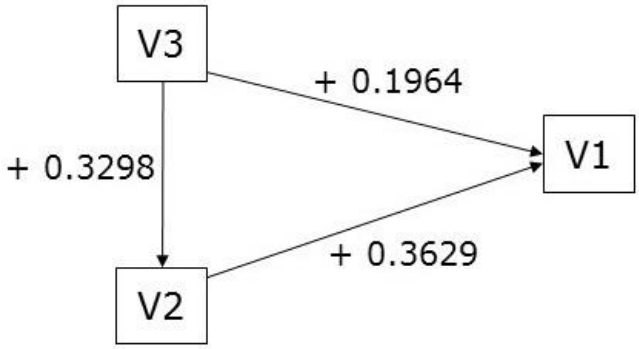
To analyse the effects of Country of origin (V3) and Work Sector (V2) on the Sharing Propensity Index (V1) we tested two structural models[15]. The first uses a reduced model not considering the effect of the Country of Origin on the Work Sector. This assumption does not reflect the reality because Not Italian respondents more often work at the University (tab. 15).

Tab. 15 – Distribution by Work Sector by Country of Origin

	Work Sector			
	Private Sector	*University*	*TOTAL*	
	%	*%*	*N*	*%*
Italians	54.17	45.83	48	100.00
Not Italians	20.00	80.00	35	100.00

Coding Italians as zero and Not Italians as one, working at University or in the Public Sector as one, otherwise coding 0, we obtain the model schema of the saturated model reported below (only direct effect, fig. 10), which seems to be the most appropriate and not reducible model as seen above. All the effects are significant at alpha=5%, also the indirect effect of the Country of Origin that influences the Sharing Propensity via the Work Sector, which is positive. The total effect of V3 (Country) over V1 (Data Sharing Propensity) is given by 0.1964+(0.3629*0.3298) that equals 0.3161. In other words, in this sample Not Italians more often work at University, so they add to their own higher sharing propensity also the fact that they work at University. We know that working at University has its own positive effect on data sharing propensity too, as the Model shows.

Fig. 10 – Path model with sharing propensity as dependent variable (V1) and Work Sector (V2) and Country of Origin (V3)



Looking for a Target to sell Data Sharing Services

The data provides some other evidence that should be outlined before answering the question *'what kind of target we are looking for*?'
The previous paragraph describes a model where the dependent variable is the *propensity to share data*. Now we observe the relations between the same independent variables with the use of metadata standard documentation, taken as a proxy of the real use of data sharing procedures. Table 16 shows that *not Italians* use metadata three times more than Italians.

Tab. 16 – Use of metadata standard by country of origin

	Using metadata standard		*Total*	
	No	*Yes*		
	%	*%*	*N*	*%*
Italians	70.83	29.17	48	100.00
Not Italians	25.71	74.29	35	100.00

Table 17 shows that those working in the public sector use metadata more than the others, but the relation is not significant at the alpha level of .1.

Tab. 17 – Use of metadata standard by Work Sector

	Using metadata standard		*Total*	
	No	*Yes*		
	%	*%*	*N*	*%*
Private Sector	48.25	51.75	33	100.00
University	32.90	67.10	50	100.00

If we look at the relationship between the use of metadata and the typology, we discover that it is significant (P value less than 1%) and that 25 people (over 83) are in the previously defined condition of *Reducible Reluctant* (tab. 18), while 20 of them (80%) are not using metadata. (Tab. 19)

Tab. 18 – Typology by use of metadata standard

	TYPOLOGY				*Total*	
	Irreducible Reluctant	*Reducible Reluctant*	*Premium Follower*	*Problematic Follower*		
	%	*%*	*%*	*%*	*N*	*%*
No	8.61	49.97	19.99	21.42	43	100.00
Yes	25.44	6.80	40.63	27.12	40	100.00

It seems reasonable to consider a target for a 'marketing campaign' aimed at promoting data sharing among those not familiar with meta documentation tools (tab. 19), having problems with data, working more often than others in a private environment (tab. 20) and more often in Italy (tab. 21, *remember that in this sample Italy stands for any place where data sharing is not sufficiently widespread*). That is the *Reducible Reluctants*, as we call them above.
First, all tables show high level of significance. Furthermore, we note that:

1. *Irreducible reluctant* (low sharing data propensity-low sharing problems), *Premium Followers* (high sharing data propensity-low sharing problems) *and Problematic Followers* (high sharing data propensity-high sharing problems) all use metadata standard more than *Reducible Reluctant* (low sharing propensity-high sharing problems). Worthy of note is the fact that, among *Irreducible Reluctant,* 9 respondents over 12 use metadata standardized within their organization, which is a limited standardization.
2. The *Followers* more often come from University. Among the *Premium Followers* academics are 10 times more than non-academics.
3. In each typology category, Not Italians are more present and in the special case of *Premium Followers* they are eleven times more than Italians.
4. *Reducible Reluctant* are proportionally more among those not using metadata standard and among *Italians*. They work at University more than *Irreducible Reluctant*.

Tab. 19 – Use of metadata standard by typology

	No	*Yes*
	%	*%*
Irreducible Reluctant	16.77	83.23
Reducible Reluctant	81.38	18.62
Premium Follower	22.65	77.35
Problematic Follower	31.99	68.01

Tab. 20 – Work sector by typology

	Private Sector	*University*
	%	*%*
Irreducible Reluctant	58.39	41.61
Reducible Reluctant	37.32	62.68
Premium Follower	8.07	91.93
Problematic Follower	25.56	74.44

Tab. 21 – Country of Origin by typology

	Italians	*Not Italians*
	%	*%*
Irreducible Reluctant	33.55	66.45
Reducible Reluctant	44.42	55.58
Premium Follower	9.76	90.24
Problematic Follower	23.58	76.42

Joining these variables with some personal characteristics, reducing to two the characters of typology (1 the *Reducible Reluctant*, 0 the others) and reversing the model taking as independent variables the work sector (0 private sector, 1 University and public sector), country of origin (0 Italians, 1 not Italians), the use of metadata standards (0 do not use, 1 use), gender (1 Male, 2 Female), Age (1=18-30 … 5=over 65), academic qualifications (1 High School, 2 Graduation, 3 Phd, 4 Master) we built a score using *Reducible Reluctant* as a target variable. The logistic model, which reaches a rescaled R square of 45% and is significant at an alpha level of 1%, retains as significant effects of independent variables at alpha level = .1 (tab. 22) the use of metadata standard (reducing the probability to be a target), the country of origin (reducing also the probability to be a target, i.e being Italian increases the probability to be a target), gender (females increase the probability to be a target), age (being young increases the probability to be a target), qualifications (higher academic qualifications increases the probability to be target too). In other words, the model says that if we do not use metadata standard, we work in Italy, we are female, young and have a higher academic qualification, the probability to be a target is higher. Although the effect of the

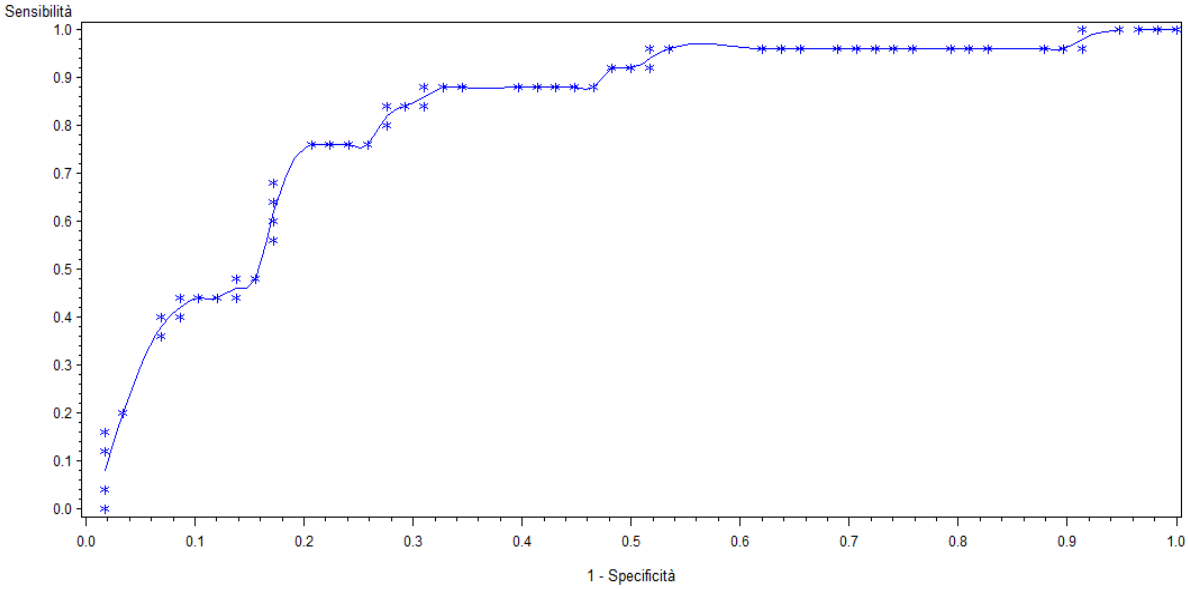
work sector is not significant, we may argue that the probability to be a target is higher also if people work in the private sector (tab. 22). *It seems that the model takes care of the warnings noted above.*

Tab. 22 – Logistic model: Analysis of Maximum Likelihood Estimates

Parameter	*DF*	*Estimate*	*Standard Error*	*Wald Chi-Square*	*Pr > ChiSq*
Intercept	1	-2.6399	2.1437	1.5165	0.2182
S05 - Use of metadata standard	1	-2.4485	0.7556	10.5000	0.0012
RA01 – Work Sector	1	-0.1203	0.8031	0.0224	0.8809
RPROV – Country of Origin	1	-2.1824	1.1335	3.7069	0.0542
F01 - Gender	1	1.3062	0.7345	3.1622	0.0754
F02 - Age	1	-0.5084	0.3108	2.6760	0.1019
RF03 - Studies	1	1.0625	0.6195	2.9415	0.0863

Using 0.45 as a cut-off point for the estimated target, as suggested by the ROC curve (fig. 11, .45 corresponds to .44 of sensitivity), there are 25 people in the estimated target (the same marginal distribution as the observed target).

Fig. 11 – ROC curve used to set the cut-off point for logistic model



Comparing the estimated and the observed targets between the random model and the logistic model by means of the appropriate confusion matrices (tab. 23, 24), we observe that: while a random model guesses 28% of the target, the logistic model guesses 60% of the target, which is more than double. When this model predicts the *target* it is wrong in 40% of the cases given the prediction (false positive). When this model predicts *not target* it is wrong in 17% of the cases given the prediction. In other words, if we

promote data services when the model says 'no' we have a high risk of being unsuccessful (83%). This risk is more than halved if we promote services when the model says 'yes' (40%).[16]

Tab. 23 - Confusion Matrices for the random and logistic model: Random Model

	TARGET		*TOTAL*	
	no	*yes*		
	%	*%*	*N*	*%*
Target Observed				
no	68.97	31.03	58	100.00
yes	72.00	28.00	25	100.00

Tab. 24 - Confusion Matrices for the random and logistic model: Logistic model

	TARGET		*TOTAL*	
	no	*yes*		
	%	*%*	*N*	*%*
Target Estimated				
no	82.76	17.24	58	100.00
yes	40.00	60.00	25	100.00

As an empirical view of model behaviour, once again we reverse the model to consider the conditional probability of the model's independent variables given the estimated target prediction.
So, by definition in our estimated target we have people that do not use metadata documentation tools (96%), work more often in the private sector (52%), work more often in Italy (72%), are more often female (84%), more often young (80% less than fifty years old), more often have a higher academic qualification (52%). This is a clear indication for the market strategy direction (tab. 25).

Tab. 25 – Characteristics of the estimated target

Independent model variables	*%*	*P-value Chisq*
Do not use metadata standard	96%	.0000
Work in private sector	52%	.1346
Work in Italy	72%	.0861
Are female	84%	.0000

Are young	80%	.0104
Have a higher academic qualification	52%	.8638

Some conclusions

What kind of story does this survey tell us? First, there at least two different attitudes about Data Sharing, probably coming from different data culture (the way data are used in a research project). These different views make different attitudes: one, the not Italians, more attentive to data sharing, recognising fewer problems doing it, being more satisfied. The other, the Italians, are less proactive for every aspect considered. We might tell other stories too, regarding several aspects we encountered during the work: it may not sound so good that the youngest age class shows the lowest data sharing propensity while the oldest shows the highest. It is not surprising instead that we can find the highest sharing propensity among the Public Administration and at University or among those with the highest academic qualifications. Also, not surprising is the fact that the highest evaluation of data sharing problems is among the private sector. A little bit more unexpected is that females evaluate sharing problems more than males, as we noted above. To conclude, in answer to the two questions reported at the beginning of this article:

1. *Is there a target for selling data-sharing services among data users?* If we think of the listed objective reasons and attitudes the answer is: Yes, if supported with a renewed promotion campaign. This target is composed by data users not having a great propensity for data sharing but recognising problems in data sharing (*Reducible Reluctant*). Once the problems encountered are resolved, it is likely that also the data sharing propensity will increase. The models show who are the reducible reluctant, as we have seen above: Italians (every country where Data Sharing is not so used), who does not use Metadata Standard, works in a private sector, is young and female with a higher academic qualification. This is most probably the most immediate target for data sharing promotion. That does not exclude that also the *Irreducible Reluctants* will be comprised in the target in the future. It will probably require more effort, given the characteristics of this Data User group.
2. *Are there any personal perspectives and attitudes that may influence data services diffusion?*
   Yes, as stated by the indexes: propensity, problem, satisfaction indexes and by every Item used to build them. Among the first five items more evaluated as data sharing problems by the respondents, three concern the problem of data ownership and data access regulation. One question comes to mind: will be the GDPR[17] be the right answer, although partial, to these troubles? But to answer this question we need further research…

What to do?

A hint comes first from the respondent distribution on the items regarding cooperation with a *broad group of researchers* and the opportunity *to create new datasets from shared data,* both of which scored highly (Data Sharing Propensity Scale). *Fostering cooperation among researchers and data reuse* is the highway we have to take in order to gain more *Premium Followers*[18].
This statement is not a great novelty, especially for Members of IASSIST that know the problem very well. Nevertheless, it is worthy of note because the response to those Items is significantly lower where the interviewed people are not very used to sharing data.
Furthermore, we have to broaden the scope of service promotion, moving from 'developed countries' to 'developing countries', those where data curation is less practised, to younger people, involving women in greater responsibility and more remunerative roles. How to do that is a matter that goes beyond the scope of this article, but one suggestion is to transform the self-referential meetings into open symposiums, for example moving from the usual locations (for Iassiters USA, Canada, North Europe) to, perhaps, less easy locations, such as southern Europe, Africa or Asia, and to less easy environments, outside the University, in an open public and private space.

Acknowledgements

I would like to thank Dr. Veronica Baldisserri for helping me in the questionnaire construction for the web and Susan Phillips for reviewing my English. The student Silvia Canavesio helped me in redaction of the final text and worked as a support in Statistical Analysis. A special thanks to the members of IASSIST that contributed to the question list and to everyone that participated in the survey

Annex 1: International survey results comparison

As a final step we report some statistics on items comparable with the ones published in *Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide,* quoted above (only the items present in both surveys). Remember that the American Surveys regard around 1000 people (tab. 26, 1329 Baseline survey, 1015 Follow-up survey), while this report has 83 people[19]).
The following table, which orders the Items on the absolute value of mean differences, show not so big differences, starting from 0.03 and reaching in two cases a difference greater than one. Furthermore when comparing the item means of the actual survey with the item means of the American Follow-up survey using the 99% confidence level, we cannot refute the hypotheses that both values are coming from the same population, at least for the first 7 smallest differences.

Means difference test between Actual Survey and Follow-Up Survey ordered by absolute value of the difference

Items Label	*Follow-Up Mean Americans*	*Actual survey mean*	*Differences between actual survey mean and Follow up Mean*	*Follow up Mean is in the Confidence Interval at the level of 95% of the actual survey mean*	*Follow up Mean is in the Confidence Interval at the level of 99% of the actual survey mean*
I would use other researchers' datasets if their datasets were easily accessible	4.330	4.31	-.017	**	***
Process for cataloguing / describing my data	3.520	3.48	-.042	**	***
It is appropriate to create new datasets from shared data	4.230	4.32	0.088	**	***
Process for analyzing my data	3.940	3.85	-.095	**	***
I would be willing to share data across a broad group of researchers	4.390	4.54	0.151		***
Process for searching for my own data	3.430	3.61	0.181	**	***
I would be willing to place all of my data into a central data repository with no restrictions	3.230	3.50	0.270		***

Items Label	*Follow-Up Mean Americans*	*Actual survey mean*	*Differences between actual survey mean and Follow up Mean*	*Follow up Mean is in the Confidence Interval at the level of 95% of the actual survey mean*	*Follow up Mean is in the Confidence Interval at the level of 99% of the actual survey mean*
Lack of access to data generated by other researchers or institutions is a major impediment to progress in science	3.990	4.27	0.275		
Tools for preparing my documentation	3.110	3.41	0.304		
Process for collecting my research data	4.050	3.68	-.375		
I am satisfied with my ability to integrate data from disparate sources to address research questions	3.190	3.58	0.390		
Lack of access to data generated by other researchers or institutions has restricted my ability to answer scientific questions	3.360	3.78	0.416		
Others can access my data easily	3.150	2.60	-.553		
Tools for preparing metadata	2.870	3.44	0.571		
Process for storing my data	3.030	3.65	0.619		
Data may be misinterpreted	4.120	2.44	-1.68		
Data may be used in other ways than intended.	4.210	1.86	-2.35		

Annex 2: IMDSM2017, Iassit Members Data Sharing Mail, FALL 2017

Iassit Members Data Sharing Mail

I would share my data but…

Members of IASSIST	Suggested text
A	My university holds ownership and won't let me
	My PI/Collaborators won't share
	I'm not done with it yet (15 years later -- haven't touched it in 14.5 years)
B	… it is difficult to compile various parts of data into a coherent dataset suitable for reuse
	… file formats are old or damaged, and there is not enough documentation (metadata) to be sure what to share
	… it is old, it doesn't answer to the questions researchers ask today
	… there is no funding to produce a reusable copy of the data
	… there are no clear rules or recommendations on which datasets should be shared
	… data is classified or a NDA was required by the funder or a partner (esp. a company)
	... but I do not know how or why (nobody has taught me that).
	… but there isn't (enough) scientific merit in or professional incentive for sharing data.
C	Why would anyone be interested in my data?
	My data are not of interest or use to anyone else.
	I have not got time or money to prepare data for sharing.
	Data sharing makes it harder to recruit participants.
	If I ask my respondents for consent to share their data then they will not agree to participate in the study.
	I don't mind making it open, but I worry someone else might object.
	People might misuse my data.
	We want people to come direct to us so we know why they want the data.
	I don't want other people taking credit for my work.

	I will if I can have an embargo…is 30 years OK?
	I want to publish my work before anyone else sees it.
	No way! My data on public attitudes towards the weather is incredibly sensitive and potentially disclosive.
	Some of what you asked for is confidential.
	My data have been gathered under complete assurances of confidentiality.
	We're worried about the Data Protection Act.
	I have collected audio-visual data and cannot anonymise them; therefore, I cannot share these data.
	I am doing quantitative research and this combination of my variables discloses participants' identity.
	My data collection contains data which I have purchased and it cannot be made public.
	There is intellectual property in the data.
	That data is already published via (external organisation X)
D	Concerns about opening up data, and responses which have proved effective
	There's no API to that system
	We're worried about the Data Protection Act (UK Law)
	I don't mind making it open, but I worry someone else might object
	It changes too quickly
	There's already a project in progress which sounds similar
	Some of what you asked for is confidential
	We don't have that data
	That data is already published via (external organisation X)
	We can't provide that dataset because one part is not possible
	What if something breaks and the open version becomes out of date?
	What if we want to sell access to this data?
	Setting a dangerous precedent

	Fraudsters use data against us
E	I promised I would keep data locked in my office
	I don't want it to be used to reverse social policies my research has been used to support
	I want to be able to co-author all publications
F	... but I have better things to do than fill out endless metadata fields that I already filled out elsewhere.
	... but I don't understand the fine print of your licence /agreement and I'm not a lawyer.
	... but your clunky system (repository) makes me lose the will to live, never mind finish my deposit.
	... but apparently even if I anonymise it there is still risk that individuals will be identified and armed, so I'll just think about it a while longer.
	... but you haven't made the case to me that it will affect my academic career one iota.
	... but nobody else in my department is doing it and really why should I be the first.
G	People will criticize my methods.
	My data is something special I can offer my own students.
	I spent a lot of money on this research, and it's too valuable to just give away.
	What if another researcher misrepresents what I did?
	What if another researcher uses my research for commercial purposes? That's not how my funder wanted this used.
H	Someone may scoop me and find something interesting in it before I have a chance to publish it!
	I'm in a niche field. Nobody else could possibly be interested in my data.
	Documenting data so someone else can understand it is complicated. Who has time?
	My institution doesn't have a repository. I don't have anywhere to share it.
	It's so confusing! I don't know where to start.
	Someone may scoop me and find something interesting in it before I have a chance to publish it!
B, D	… legal reasons prevent it (for example because of personal data act) or research ethics code prevents sharing
B, E	… I've promised to my research subjects that I don't share the data or I haven't said anything about archiving (no informed consent)
B, G	… sharing might lead to need to give advice/guidance to those who reuse of the data (no time for it)

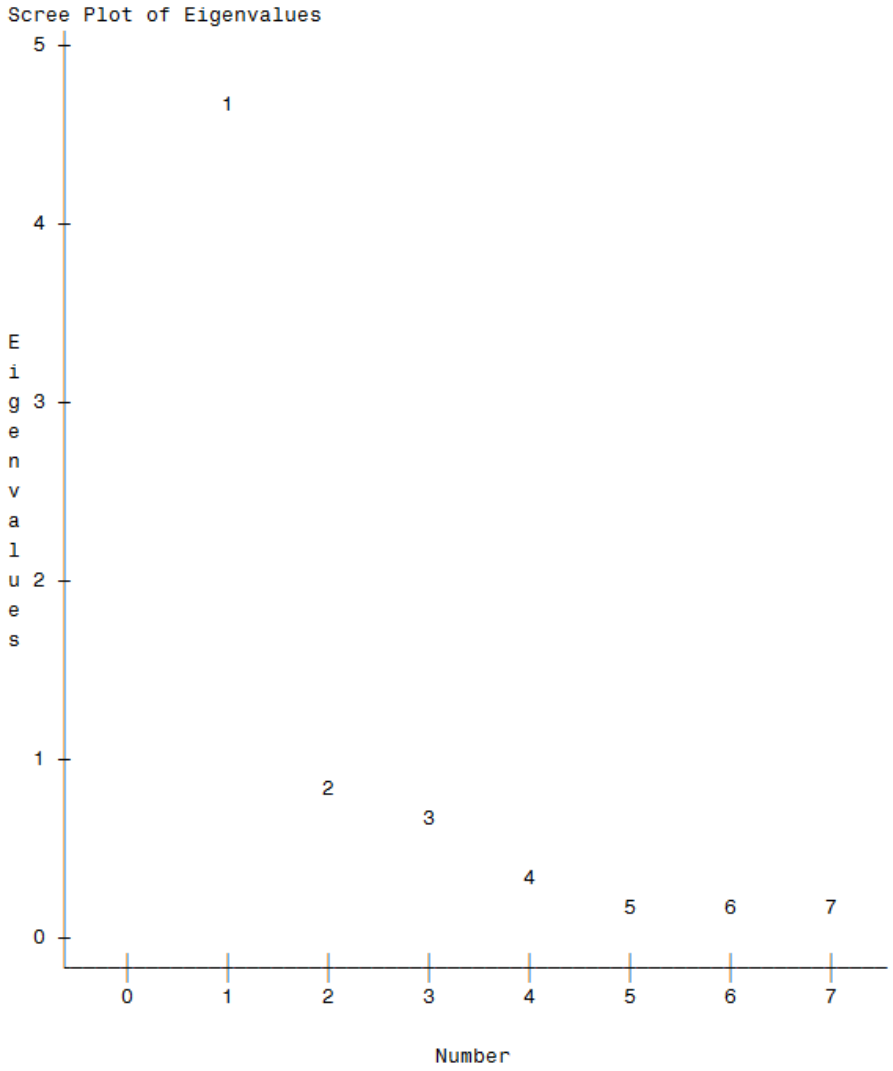
B, H	… it would be a copyright infringement or I cannot make the copyright clearance
A, B, E	… no one else can understand the data, it is too personal
B, C, E	… it is customary to delete the raw data after the research has been carried out
B, C, H	… data security issues regarding identifiers prevent it; data cannot be anonymized or it is too expensive to do, or it becomes useless when anonymized
B, C, D, I	… there are no clear definitions of ownership or usage rights
C, D	We can't see the benefit.
	If we publish this data, people might sue us.
	Terrorists might use the data.
	We'll get spam.
	It's too big.
C, H	It's my data. I don't want to share it, and that's all there is to it.
	Other researchers would not understand my data at all – or may use them for the wrong purpose.
C, D, I	People will contact me to ask about stuff
	People will misinterpret the data
	My data is not very interesting
	I might want to use it in a research paper
	My data is too complicated.
	My data is embarrassingly bad
	It's not a priority and I'm busy
C, D, H	I don't own the data, so can't give you permission.

Annex 3: Factor analysis for work satisfaction index, problems importance index and sharing propensity index

For each factor analysis we report the factor variance (eigenvalues) scree plot and the factor pattern matrix. The scree plots report the extracted factors (components) ordered by the quantity of explained variance of the components extracted. In every case we observe that the first factor explains much more variance than the others, which will be the selected component for further analysis. The factor pattern matrix tells us in a synthetic way which of the original items counts more in the computation of the factor final value helping us to give a meaningful sense to the extracted factor.

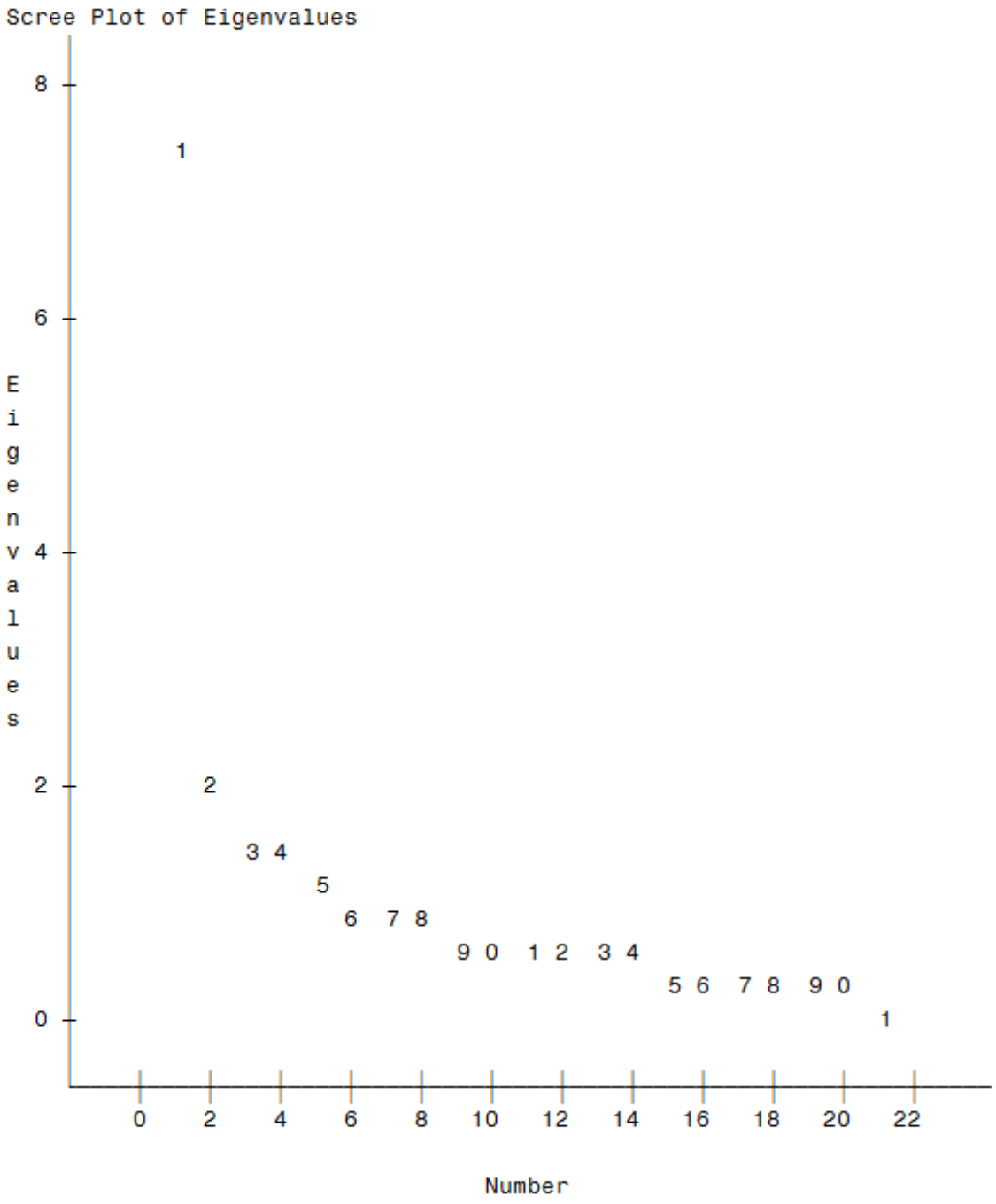
From these analyses we can say first that the factor extracted is for every analysis the most important factor (in terms of explained variance). Second, we can get a more precise Idea of the meaning of each factor. So, for the work satisfaction index we can say that all items used have a similar importance on the final factor. Just the item “tools for preparing metadata” is slightly less important. The problems importance index summarizes several items. Therefore, we see more items that seem to be less important in the computation of the factor. Among them we find: “My data have been gathered under complete assurances of confidentiality”, “Data belong to my organization/university/company and it doesn’t give me the permission” and “I have not got time to prepare data for sharing”. For the sharing propensity index the item “It is appropriate to create new datasets from shared data” shows a slightly smaller correlation coefficient with the respective factor values than the others.

Factor analysis for work satisfaction index



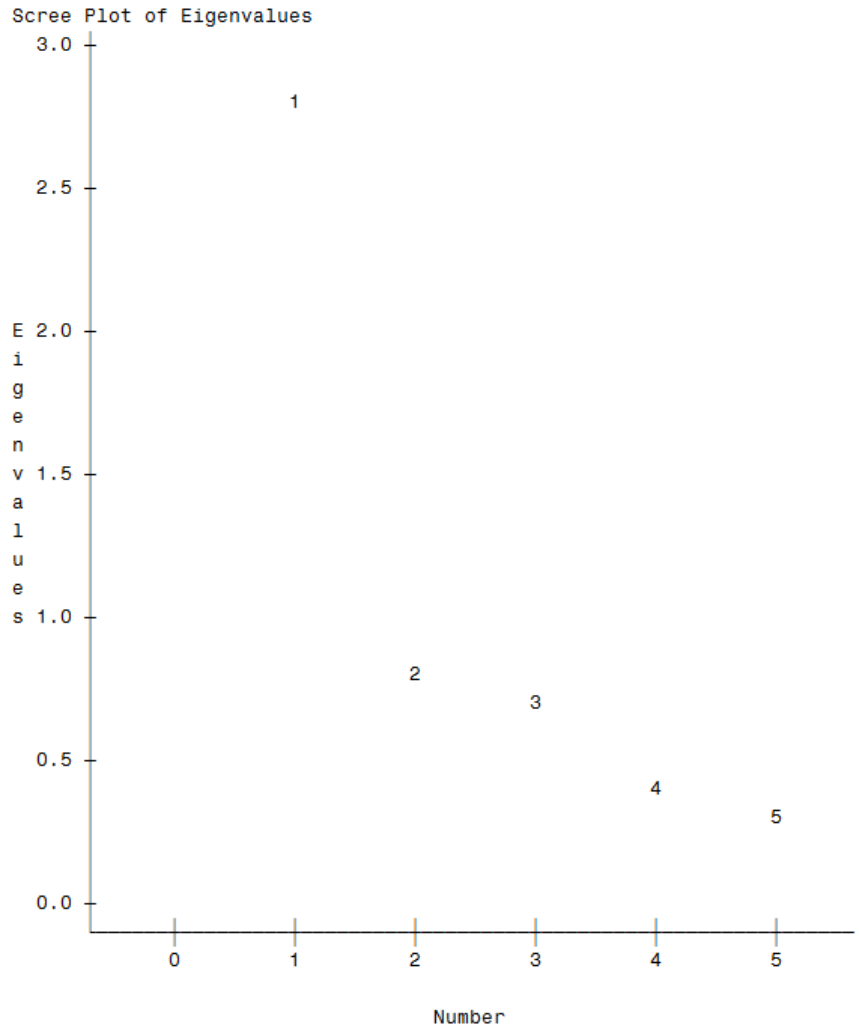
Factor Pattern	
	WSI - Factor
Process for collecting my research data	0.81748
Process for cataloging / describing my data	0.84129
Process for storing my data	0.87918
Process for searching for my own data	0.87144
Process for analyzing my data	0.75725
Tools for preparing metadata	0.65962
Tools for preparing my documentation	0.83165

Factor analysis for work problems importance index



Factor Pattern	
	SPI-Factor
My data have been gathered under complete assurances of confidentiality	0.38103
There are no clear definitions of ownership or usage rights	0.65201
There is intellectual property in the data	0.66610
Data cannot be anonymized	0.52725
I have collected audio-visual data and cannot anonymize them	0.44449
Data belong to my organization/university/company and it doesn't give me the permission	0.39950
My data change too quickly	0.66832
Lack of funding	0.46256
I have not got time to prepare data for sharing	0.40837
Files are old or damaged, and there is not enough documentation (metadata)	0.57591
We no longer have datasets	0.55692
I would not know where and how to share data	0.53569
My data are old, they don't answer to the questions researchers ask today	0.67454
I can't see the benefit	0.65204
People might misuse my data	0.67071
I spent a lot of money on this research, and it is not economically convenient to share it	0.66127
It could affect negatively my career	0.68575
It's my data. I don't want to share it	0.65698
I don't want other people taking credit for my work	0.64718
I would lose control of the data	0.69237
I have not finished analyzing the data yet	0.60151

Factor analysis for work data sharing propensity index



*

Factor Pattern	
	SPN-Factor
I would use other researchers' datasets if their datasets were easily accessible	0.74303
I would share my data if someone took care of the archiving process	0.74334
I would be willing to place all of my data into a central data repository with no restrictions	0.79848
I would be willing to share data across a broad group of researchers	0.77074
It is appropriate to create new datasets from shared data	0.66703

References

Bonifacio, Flavio (2017) 'Working Across Boundaries – Public and Private Domains', Part 3 – A follow-up Survey, (Available at http://doi.org/10.5281/zenodo.1120237)

Doorn, Peter and Tjalsma, Heiko (2007) 'Introduction: archiving research data', Springer Science+Business Media B.V.

Hatcher, Larry (1994) 'SAS System Factor Analysis and Structural equation modelling', SAS Institute

Horton, Laurence (2016) 'LSE Research Data Management Data Sharing Objections FAQs and Naughts and Crosses Game', (Available at http://doi.org/10.5281/zenodo.61978)

Kim, Youngseek and M. Stanton, Jeffrey (2012) 'Institutional and Individual Influences on Scientists' Data Sharing Practices', Journal of Computational Science Education, Volume 3, Issue 1

Noble, Susan; Russel, Celia and Wiseman, Richard (2012) 'Mind the Gap: Global Data Sharing', IASSIST Quarterly, Vol. 35, n° 3

Qualtrics (2010) 'The 1936 Election – A Polling Catastrophe', (Available at https://www.qualtrics.com/blog/the-1936-election-a-polling-catastrophe)

Rasmussen, Karsten Boye (2014) 'Social Science Metadata and the Foundations of the DDI', Vol. 37, n° 1

Ribeiro, Cristina and Matos Fernandes, Maria Eugenia (2012) 'Data Curation at U. Porto: Identifying current practices across disciplinary domains', IASSIST Quarterly, Vol. 35, n° 4

Tenopir, Carol; D. Dalton, Elizabeth; Allard, Suzie; Frame, Mike; Pjesivac, Ivanka; Birch, Ben; Pollock, Danielle and Dorsett, Kristina (2015) 'Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide', (Available at https://doi.org/10.1371/journal.pone.0134826)

Yang, Meng-Li (2013) 'Strategies of Promoting the Use of Survey Research Data Archive', IASSIST Quarterly, Vol. 36, n° 1

Notes

[1]Flavio Bonifacio, Metis Ricerche srl, Via Camerana 6, I-10128 Torino, Italy. Contact e-mail: flavio.bonifacio@metis-ricerche.it.

[2] European Data Definition Initiative, 17th Congress, Lausanne, 5-6 December 2017. We included this file of emails to increase the number of respondents. There is no doubt that the participants at EDDI17 are data users.

[3] Tenopir, Carol and others, op. cit.

[4]Tenopir, Carol; D. Dalton, Elizabeth; Allard, Suzie; Frame, Mike; Pjesivac, Ivanka; Birch, Ben, et al. (2015)

[5]Kim, Youngseek and M. Stanton, Jeffrey, (2012), Doorn, Peter and Tjalsma, Heiko, (2007),Noble, Susan; Russel, Celia and Wiseman, Richard, (2012),Ribeiro, Cristina ad Matos Fernandes, Maria Eugenia, (2012) Yang, Meng-Li, (2013),Rasmussen, Karsten Boye, (2014)

[6] See http://doi.org/10.5281/zenodo.61978 quoted by Horton, Laurence, IASSIST MEMBERS DATA SHARING MAIL

[7] See IASSIST MEMBERS DATA SHARING MAIL, FALL 2017 (IMDSM2017 in the following)

[8] The sample is not random because only those who want to respond to a questionnaire actually answer. The Italian sample list was extracted from our mailing list and from participants at the Turin Seminar amounting to a total of about 500 people (525). The international sample list comes from the Iassist Member Directory (431 people) and from the participants at the EDDI17 congress (69 people), resulting in a total of 500 people (percentage of respondents: 9% and 7% respectively).

[9] Following the suggestions of the reliability analysis, for the last scale we excluded item 4, causing a decrement of Cronbach's alpha.

[10] The indexes have been calculated using PCA, rescaled with range *0-100* and recoded in three classes. The classes limits, expressed in standard units, are: 1, less than *0.431*; 2, between *-0.431* and *+0.431*; 3, greater than *+0.431.* In the case of normal distributions, the classes would be uniformly distributed. We have also recoded them into two categories, less or greater than the mean value

[11] Which is zero for the standardized indexes

[12] First category includes the respondents with both indexes below the mean value, second category the ones with a propensity index below the mean value and sharing problems index over the mean (that means *few sharing problems*), etc.

[13] Definition: if Data Sharing were a subscriber of FB, then those that follow it are Followers

[14] Only in the logistic model presented below do we use more variables as independent variables

[15] We follow the notation proposed by L. Hatcher, (1994)

[16] We consider 'at risk' when we promote services to the wrong target. This happens every time the model predicts correctly 'not target' and predicts incorrectly 'target'. This analysis is not complete because we do not have enough cases to test the model.

[17] GDPR (General Data Protection Regulation) regards the management of personal data. Personal data here means: "Any information relating to a data subject. Sensitive personal data, which attracts a high degree of protection, is data which is in relation to race, political opinions, health, sexual life, religious and other similar belief, trade union membership and/or criminal records." See the link https://www.audiencedatasharing.org/legal-information

[18] Bonifacio, Flavio (2017)

[19] It is well known that the sample dimension is not the only thing that influences the outcomes of a survey. An example is the Poll conducted in 1936 in USA about the Presidential Election, when the Poll using 2 million surveyed persons predicted A. Landon as winner against F.D. Roosevelt https://www.qualtrics.com/blog/the-1936-election-a-polling-catastrophe/