Information **2011**, *2*, 455-459; doi:10.3390/info2030455

OPEN ACCESS

information

ISSN 2078-2489

www.mdpi.com/journal/information

Article

On Symmetries and the Language of Information

György Darvas [1,2]

[1] Symmetrion

[2] Institute for Research Organization, Hungarian Academy of Sciences, 18 Nádor Street, Budapest H-1051, Hungary; E-Mail: darvasg@iif.hu

Received: 19 May 2011; in revised form: 28 June 2011 / Accepted: 19 July 2011 / Published: 22 July 2011

Abstract: Many writings on information mix information on a given system (I_S), measurable information content of a given system (I_M), and the (also measurable) information content that we communicate among us on a given system (I_C). They belong to different levels and different aspects of information. The first (I_S) involves everything that one possibly can, at least potentially, know about a system, but will never learn completely. The second (I_M) contains quantitative data that one really learns about a system. The third (I_C) relates rather to the language (including mathematical) by which we transmit information on the system to one another, rather than to the system itself. The information content of a system (I_M—this is what we generally mean by information) may include all (relevant) data on each element of the system. However, we can reduce the quantity of information we need to mediate to each other (I_C), if we refer to certain symmetry principles or natural laws which the elements of the given system correspond to. Instead of listing the data for all elements separately, even in a not very extreme case, we can give a short mathematical formula that informs about the data of the individual elements of the system. This abbreviated form of information delivery includes several conventions. These conventions are protocols that we have learnt before, and do not need to be repeated each time in the given community. These conventions include the knowledge that the scientific community accumulated earlier when discovered and formulated the symmetry principle or the law of nature, the language in which those regularities were discovered and formulated, for example, the symmetry principle or the law of nature, the language in which those regularities were formulated and then accepted by the community, and the mathematical marks and abbreviations that are known only for the members of the given scientific community. We do not need to repeat the rules of the convention each time, because the conveyed information includes them, and it is there in our minds behind our communicated

data on the information content. I demonstrate this by using two examples, Kepler's laws, and the law of correspondence between the DNA codons' triplet structure and the individual amino acids which they encode. The information content of the language by which we communicate the obtained information cannot be identified with the information content of the system that we want to characterize, and moreover, it does not include all the possible information that we could potentially learn about the system. Symmetry principles and natural laws may reduce the information we need to communicate about a system, but we must keep in mind the conventions that we have learnt about the abbreviating mechanism of those principles, laws, and mathematical descriptions.

Keywords: symmetry; language; information reduction; measure of information

1. Symmetries and Reduction of Information

Laws of nature, according to the concept established in the early 17th *c.*, determine the state of an object or a system at an earlier or later moment expressed in an abbreviated form. This means, we do not need to execute measurements and record its results in every moment to get information about the state of a given object or system. We know those states predicted by the appropriate law of nature. Tycho Brache recorded many data on the orbital (spatial) position of the planets in many different consecutive (temporal) moments. Based on these data, Johannes Kepler formulated his celestial laws in short mathematical formulas. These laws determine the position of the planets at any moment in time without the need to measure their position in the sky. We can obtain information on the observable place of a planet at any given time by the help of Kepler's laws. The three Kepler formulas contain infinitely more information on the position of the planets than all data collected by Brache and his predecessors for many years, but we can nonetheless communicate that information in a very short form.

The same can be said regarding other laws of nature. They can be tested, and can be applied at any place and any time. Kepler's Laws, e.g., hold not only in Prague, where he established them, and not only in the first years of the 17th century but rather, they are valid anywhere and at any time. Scientific laws express a symmetry of nature, namely, they hold if we execute a shift in space and/or in time, they conserve their validity. They are invariant under spatial and temporal translations. Some laws are conserved under other, further changes. For example, laws of motion observed by Galileo are valid under a switch to another reference frame moving with a fixed (not too high) velocity. This invariance under change of inertial systems was extended to any velocity by the Lorentz transformation, and to accelerating reference frames by the General Theory of Relativity.

In biology, we can give the genetic code as a sequence of the four nucleotide bases in the RNA or DNA of an individual organism. This sequence determines, among others, the proteins of the given living being. We know the law that any triplet of the nucleobases (codon) determines one of the twenty essential amino acids used by living cells to encode proteins. Thus, we can reduce the data (by two thirds) to give the information for the proteins. It is enough to describe the sequence of the triplets which encode each one of the twenty amino acids building the structure of the individual protein.

A full series of natural laws and their applications could be listed as examples of how the laws may reduce the data that we need to give to get all the necessary information one needs to know about a system. All these laws serve to reduce the information one needs to communicate to others who want to be informed about the states of a system.

The laws do not determine the total information content of a system. Other laws give information in another abbreviated form on other properties of the system. Further, we do not know every law that governs the states or the behavior of a system, so there always remains information that must be given individually at any moment.

At the same time, laws of nature (in the above mentioned meaning) are not the only symmetries that allow us to give information about a system in an abbreviated form. For example, we can give information about a picture to assign three color codes to each of its pixels in all positions. If we know that the picture is mirror symmetric in respect to one of its axes, we do not need to repeat assigning the codes to all pixels. The information on its mirror symmetry allows us to give only the half of the color codes, all the rest are given in the abbreviated form "mirror symmetric in respect to ...". Similar reduction of the communicated information can be executed in the case of "fivefold symmetric in respect to a perpendicular axis ..." (to one fifth), or in the case of "translational symmetric with a periodicity of ... in the direction(s) ..." (to—potentially—infinitely multiple times), *etc.* Symmetries can function to reduce the information on a system, similar to the role of the laws of nature. This function can be fulfilled not only by geometric symmetries. Color symmetry, tonality in music, electric charge conjugation and generalized abstract charges in physics, *etc.* can fulfill the role of reducing the necessary communicated information as well. (For example, we can say that the abstract charges of the strong physical interaction, nicknamed by physicists abstract "colors", belong to the symmetry of the $SU(3)$ group, instead of listing all the eight possible ways by which the three "color charges" can be transformed into each other). However, symmetries do not concern all information on a system either.

2. Language of Information

Information in the above section was used in at least three different contexts. Many writings on information mix them.

Information on a given system (I_S) involves everything that one possibly can, at least potentially, know about a system, but will never learn completely. We could say, I_S incorporates the full information content of a system.

The measurable information content of a given system (I_M) contains quantitative data that one really learns about a system. The I_M information content of a system (what we generally mean by information) may include all (relevant) data on each element of the system. Relevant data mean data that we can really obtain from a system.

The third type is the (also measurable) information content that we communicate (I_C) amongst us about a given system. It relates rather to the language (including mathematical) by which we transmit information on the system to one another, rather than to the system itself.

The three defined kinds belong to different levels and different aspects of information.

As we saw in the previous section, we can reduce the quantity of information we need to mediate to one another (I_C), if we refer to certain symmetry principles or natural laws that the elements of the

given system correspond to. Instead of listing the data for all elements separately, even in a not very extreme case, we can give a short mathematical formula that informs us about the data of the individual elements of the system. This abbreviated form of information delivery includes several conventions. These conventions are protocols that we have learnt before, and do not need to repeat each time in the given community.

These protocols include knowledge that the scientific community accumulated earlier. The scientific community accumulated this knowledge when discovered and formulated the actually applied symmetry principle or the given law of nature. This knowledge also includes the language in which those regularities were formulated and then accepted by the community, and the mathematical marks and abbreviations that are familiar only for the members of the given scientific community. We do not need to repeat the rules of the convention each time, because the conveyed information includes them, and it is there in our minds behind our communicated data on the information content.

As shown above regarding the example from mirror symmetry, if we say that the elements of a system are placed mirror-symmetrically, we do not need to give the position of all elements individually. It will be enough to give the positions of half of the elements and add that the rest is placed mirror-symmetrically. Of course, this statement includes all preliminary information on what we mean by "mirror-symmetry".

Quantitative laws (like those of Kepler) make it possible to use much less information (I_C) to obtain and mediate the measurable information (I_M, in Kepler's case, on the position of the planets). However, over the few mathematical variables that appear in the mathematical equations that express a law of nature for the scientists (for example, those, applied in the Kepler formulas), the abbreviated information includes all the intellectual work invested by the scholars who discovered the laws. One should add the information represented by those.For example, in the case of the Kepler's laws, the three equations include all experimental data, as input information, observed and collected by Brache and his predecessors, and all mathematical knowledge by which we understand what those formulas mean and how to use them.

Another example, cited above, is the information content of a DNA. We can give it as a sequence of the four nucleotide bases. Their sequence determines, among others, the proteins of the individual organisms. In consequence of what we learnt, the law that any triplet of the nucleobases (codon) determines one of the twenty essential amino acids used by living cells to encode proteins, we could reduce the data by which one can give the information for the proteins, by the means of giving the sequence of the triplets (and not all nucleobases) that each encode one of the twenty amino acids. However, in this case too, our knowledge lies in the background, concerning the knowledge about the law of correspondence between the codons' triplet structure and the individual amino acids that they encode, and all our knowledge about how proteins are built of amino acids.

How do we measure these three kinds of information?

The information content of a system I_S is an idealized, unreachable quantity. It can be described by means of the language of the Shannon type entropy, what is an analogy borrowed from thermodynamics.

We communicate the information on a system I_C among us either in different semantic languages, or by the means of comparable (for example, binary) codes (sequence of signs 0 and 1). (Languages are not the best, most appropriate ways to measure the information that they carry. For example, what

we define as the symbol "20" is labeled in English by "twenty", in German by "Zwanzig", in French by "vingt", and so on, not to mention the non-Latin-based character languages. Several (para-scientific) works, for another example, attribute meaning to the number of characters of different phrases in the Bible, nevertheless, if one accepts the validity of their results, it remains questionable whether it holds for the quantitative information in the language version (Greek, Hebrew/Arami, Latin) that the given author analyzed. The abbreviated information that we obtained as a result of reduction by means of symmetry principles (including scientific laws, as well as their mathematical symbol system) is measured in terms of I_C, and is meant in coded terms. (These coded terms may be binary, hexadecimal from among the widespread coding systems, or any other).

The measurable information I_M that can be obtained from a system is generally described by binary codes. However, this binary information should be brought into correspondence with the measure of information expressed in thermodynamic terms. Here we should take into account that I_M (in contrast to I_S) corresponds rather to the concept of extropy (the distance of a physical system from its equilibrium state on the entropic scale, *i.e.*, the distance from the maximum of the entropy) [1], than to that of entropy, if one takes into account the thermodynamic analogies.

3. Conclusion

The information content of the language by which we communicate obtained information, cannot be identified with the information content of the system that we want to characterize, and moreover, it does not include all the possible information that we could potentially learn about the system. Symmetry principles and natural laws may reduce the information we need to communicate about a system, but we must keep in mind, and add the information content of the conventions that we have learnt about the abbreviating mechanism of those principles, laws, and mathematical descriptions.

Acknowledgements

Instead of a detailed reference list, I express my thanks to the many contributors of the [FIS] discussion list who contributed to developing and shaping the idea presented in this paper, which could reflect only a few aspects of everything that arose in the discussion on language, cognitive science and information. The ideas expressed in this paper were inspired by their comments in the discussion. References to classics, like the Kepler formulas, the DNA nucleotide base sequences, Shannon entropy, *etc.* are commonplace.

Reference

1. Gaveau, B.; Martinás, K.; Moreau, M.; Tóth, J. Entropy, extropy and information potential in stochastic systems far from equilibrium. *Phys. A* **2002**, *305*, 445-466. Available at: http://martinas.web.elte.hu/michel.pdf (accessed on 22 July 2011).