**Tools and Technologies of Language Documentation**

**Prof. Bornini Lahiri and Prof. Dripta Piplai (Mondal)**

**Department of Humanities and Social Sciences**

**IIT Kharagpur**

**Week-04**

**Lecture-16**

Lecture 16 : Metadata & Other related information

Welcome to the 16th lecture of the course, Tools and Technologies of Language Documentation! Today, I will talk about metadata and other related information. So, metadata is very important aspect of language documentation. It is an older concept, where we had catalogs in libraries or different types of catalog. It is made in a new digital era in a newer form, though the thing is older, yet we see new system added to it. So, we will talk about that. What are the information which are required for metadata, which we need to collect from the field itself, that I will talk about and I will also talk about details, which we require to know in the field about the language experts, details which we need to note about each of the sessions where we record.

So i will talk about all this. Along with that, I will also talk about metadata creation. How it is created after processing the data? I will talk about thick and thin metadata, utility and use of metadata and also, I will show you some models of metadata which are practiced across different projects. So, let us know what is metadata? Metadata is basically data about data.

So, when we have a huge amount of data, we need to know where we have kept each of the files. So in one file, there can be like 1000 words which you have collected from the field. So, we should know and like that, you can have say 10 files. In each of the files, you have, maybe like 100 words in one file, 100 sentences in another file, 10 folk lores in another file. So now after some time, you will need to know where you have kept this.

You need to know which file or which folder, which contains words related to body part is kept. Now, you cannot keep on searching through all the folders that will take lots of time. So, you need to have a catalog of it, that is what is metadata, where we have the list being done, where we know where we have stored this special data like where we have kept the words related to body part, where we have kept sentences related to say

agreement or negation. So, everything is arranged in a systematic manner. So, that when we need to go back to that file or folder, we can easily trace it.

Otherwise it will not be possible to know where is it, if you are in the field you have collected huge amount of data, how will you do that; that is at an individual level. Think about the huge projects. The projects where we have lots and lots of people working together and then there are huge amount of data being collected and in these projects, they are not working in one language, they are working in many languages. So there are lots of languages- 50/60 languages or even more, there are lots of people working from different parts of the world, but the data is being archived in one place, in one place they are kept safe for preservation, and also for access if one wants to access the data, one can access through that archive. Now, in that archive the data should be kept in a very very systematic order.

So, maybe one researcher is working today and another researcher wants to look at that data, maybe one wants to work in an interdisciplinary area and wants to look at the data. So, what will that person do? If it has been cataloged in a proper manner and if it is accessible then the person can access that data. So, it is very very important to arrange the data in an organized way and not only arrange it, to keep a track of it, keep a note of it, where which data has been arranged or kept. Basically that is all about metadata. So, this catalog information which of course, nowadays is a kept in digital form.

So, that is the digital resource or metadata. It supports archives and collection, it is important for management and protection of the materials. At the same time, it also helps in proper citation of the archives. So if there is some data already kept in an archive, I am getting access to it through this catalog and I am using it for certain purpose, I can also cite it that this is not my data; someone else has taken it and then, I can properly give the citation. So, for all this purpose metadata is very very important and nowadays, lots of people who are working in the field of language documentation, they are actually talking a lot about metadata creation, how it can be more systematized, how it can be more standardized.

So, there can be different projects following their own ways of cataloging, but then there can be some uniform ways where everyone is following that. So, what are the better ways? So, there are lots of discussion going on about that. So, some basic information about the data which is kept in all the metadata: these are from whom the data has been collected. So, these are just few fields which I have kept here which you can see and understand. So, basic information like from whom the data was collected, Mr.

Z or whatever. So people, at times, can give codes to the language experts from whom

the data was collected or sometimes, one can even mention the name; generally, codes are given sometimes to keep the names anonymous, from where it was collected. So maybe, you will write the place and of course, there will be more information like the name of the village, district and all, but everything might not be possible to keep in the metadata. So, that information can be kept in the larger file. So what type of data? May be narrations, equipment used, when and on which time and which date it was collected, who was the person who collected it.

It is also important, name of the main researcher, name of the other team members who were involved, maybe someone was holding the recorder, other one was asking questions, so the team members and remarks, if any. It was raining very hard and so or it was about to rain, so Mr. Z wanted to leave. So maybe in his narrations, he missed some part. So when you are collecting the same narration from someone else, that person may give you more about the story, more about the folklore, but this person did not.

And then you might come back to the remarks and know that maybe this person was in hurry, so he skipped that part. So, sometimes you can mention that the person was not well. So, maybe he could not talk properly or the person was drunk. So, there can be various reasons or various remarks which can be written for purpose which will be handy when you are actually analyzing the data. So now, some of the information which we need to collect in the field like the name of the person, from whom you are collecting the data, his or her age, address if the person has migrated from some place or not and if the person has migrated, then what was his or her previous address, those also influence one speech that is why it is important to mark those.

Gender, why gender is important? Gender is important because at times, we see in a particular language due to gender, there can be variation in the language. So, educational qualification again, influences the language. So, maybe the person is educated in an English medium school, where his or her speech was affected. Occupation. So, if the person is fisherman, then of course, he will give you more terms related to fish.

If the person is farmer, he can give you more terms related to farming and also, occupation can also be helpful and as I mentioned, there can be other remarks. This is a very very short and basic format. Of course, other fields can be added to it, like in the case of women, it can be added that this women is married from this particular community. So, that information can be there. There are inter community marriages which again influences the language within the family.

So, the children might be learning both the varieties or both the languages. So, to look into that, for that a field can be added for married women that she came from another

village. So, her language is little different and there can be other informations which can be added to it. So, once this information is already there about a particular language expert, then we do not need to collect it every time we sit with the person because basic information we have. But every time we sit, we need to have the session details because every time, in every session our session details will change.

So, session detail means description or details about the session when you started the session, what was the time, what was the date, place where you started. So, if it is taken that you are actually working within a village, a particular village of a particular district, every time you will not write the name of the village because it is taken that you are working in that particular village. It is there in the main information part or main metadata. In the 'place', you might say that in someone's house, in a temple, in a playground. So, these details can be mentioned.

Equipments which are used: So, there can be video recorder, audio recorder and not only video and audio recorder is what you mention, you also mention the model names, because that actually changes the quality of the voice or pictures or video. Any remarks and also the researcher's name, who is collecting the data because he or she is responsible for the session. So, the person who is responsible for the session is named and sometimes also, the team members are also named. So, the main person is named who was actually conducting the work and then there were others to help him or her. So, all these informations are kept in every session.

Sometimes they are done in written form, but nowadays, we see that all these are generally recorded. So, when you start your recorder, then you actually record that "Today is 6th of May, it is 9 am, I am standing in Mr. Z's house and I am here to talk to him about body part terms, I am using this particular equipment and Mr. Z seems to be little lazy today as it is a Sunday, but we will talk and we will see what we can collect". So, these type of recording, may be basic something you can collect to just have the background of the recording and then, this information afterwards is taken in the metadata .

So now, this is a sample of metadata where you can see and again, this list is not limited; there can be more fields to it. So, here you can see this is when there are 4, 5 or many languages being worked upon in a particular project. Then we make metadata for a particular language for a particular field like this, where we have language code because there are lots of languages, so language code is important. Sometimes language codes which are already there for different languages, so we see lots of language codes are already given. So, we can use them like ethnolog or other sources which has given language codes, those can be used or you can also create your own language code;

projects can create their own language code.

So, you can see this is language code, Mah is given. Then there is RRA, which means the main resource person. So again, I will like to say these short forms- RRA or RA or PH, LX these are again project specific. So this, project decides to use these shorter forms. Another project can use something else or there are some standard forms which I will talk about; those formats can also be used.

So it depends upon the project. Those who are involved in the project, what do they decide, how do they decide to create the metadata or the names of the fields. So, this RRA here means the responsible person in the field, then the equipment used and you can also see that there are short forms given. So, the whole name is not written every time because it is tiring job and also, why to write; that will again take space. So, short forms are given sometimes to the researchers and also to the language experts. Audio: the type of the recording is audio; format is given, subject is word list that is why WL.

So again, there is a short form being used. RA is the person who was there also, along with the team members. So, PH is the form, phonemic form in which the word particular word. And again, lx is written in bengali script. Gloss is the meaning of the word, and you see that this is 'sim' which means 'hen' that is D.

A. SD is semantic domain and D.A. is domestic animal, so semantic domain is domestic animal LE is the language expert. So, his name is something with S and S. Date of data collection, location of data collection, source file name. Now, source file name means when we collect a data after every word, we do not pause or stop our recording. So in one go, we might collect like 50 words.

So maybe in this main file, we have 50 words related to domestic animals, but from there, a particular word 'sim', which means 'hen' has been cut and kept as a file. So, this the segmented sound file has another name. So, source file has one name and segmented file, the sound which has been cut from the main file, has another name. So, what happens is that, when you come from the field you cut the words out of the main file and then you arrange it. So, all the words related to maybe, domestic animal or wild animal or flowers are kept in one folder with the particular file names.

So like that, we can see it and when we create these file names, they are also created in a very organized manner. So, when we see D.A. 36 that means, this is the 36th word of the semantic domain list of domestic animals. So after processing the data, these all work is done.

So, in the field what we do? In the field generally, we collect details about the language expert, we collect details of each session. But when we come back, we organize our data after processing and in the processing, it involves cutting a little bit of annotation, glossing, all these things. So after doing all these things, data is properly arranged and there we create this catalog for each and every file. It might be audio file, it might be video file, it might be image file, any file. We can create this type of meta and that is actually done, so that we can easily access it.

So the structured information, describing characteristics of events and recordings and properties of data file. So, what we see is that they are in a structured manner. The information is structured, arranged in a proper manner and where we have details about the events and the recordings and also the properties, as you saw that the file type, the format type, everything is being mentioned. So, metadata is information about resources, information about language resources, lexicons, audio tapes. So, everything language description everything can be there in metadata.

So, metadata is not the actual data, metadata is the information about the actual data, how you can go to the actual data, you do not need to actually search it anywhere. You can follow the path which metadata shows you and you can easily go through that path to reach the main data. So, metadata should have a structured unified and regular format, so that it can be easily retrieved by mechanical search engines. So, maybe you are searching for say body parts or searching for domestic animals and easily, you can get it. So in this previous slide, when you see that there are certain terms being used like SD, LX, PH or any anything that you see, all these have to be uniform across the project.

It cannot be like for one language we are using PH, but for other language we are using some other term for the same field. Or if we are using RRA for the responsible person in one language, in the same project we cannot use some other term for another language. So all these fields have particular, you can say acronyms or you can say the names given to each of the field, this is fixed. It cannot change or you cannot customize it every time you enter data.

It has to be pre-decided and it has to be fixed. Lots of thought process goes behind it before fixing it, but once fixed, it is maintained throughout the project across the languages. So it has to be uniform and that is how it works for also the search engines. So, Nathan and Austin talks about thin and thick metadata. Thin metadata is only intended to support resource discovery.

So, search engines basically. So, when I say that I want to see what do you call knows in say Kurmali and in Mahali, in Lodha, in Dhimal. I can actually write knows and search

for all these different languages and maybe I will get if one project have all the data for all these languages I will get words for different words for the same word in all these languages. So, that is how thin metadata actually works. So, it has cataloging. So in catalogue, what do we have? We have the name, the title, speaker's name, the collector's name, who have collected the data, time and place of recording and language name.

So basically, all these information you record in the field itself. So, all these things. And then there has to be some descriptive information, information about the content, relationship to other resources. So maybe information about the content means you can say that domestic animals.

As I mentioned semantic domain. You can also mention parts of speech, like noun and relationship to other resources. So, if I am saying domestic animal, maybe it is a sub-domain of the main domain, animal. So I will write that relation that it is a sub-domain of the main domain, animal. So within the animal semantic domain, we can have several sub-domains like domestic animal, wild animal, reptiles, whatever we want. Then there has to be structural information, what structural devices and patterns exist in the document.

So, how it has been structured? So, maybe if it is gloss then or meaning. So, what are there? So, we can say that gloss is there. So, we give the meaning. If there are folklores, then we can say that the translation is there, transcription is there. Then the technical information: performance and preservation information, description of format.

So how it was actually recorded, what is the format, what are the devices used to record it, all these information and of course, description about it. So, we can write the size also, we can write the format, all these technical information. And then comes the administrative information which is also very important, which is of course, decided by the project member; its not decided by a single person generally. So, what happens is that administrative information means who can access the data.

There can be some data which you cannot make public. So, there has to be some restrictions or copyright issues, who have the copyright. So, all these type of informations are kept in the administrative information. So, administrative information also tells how one can access the data. Some data can be freely available, some data there might be some copyright issues. So, how it can be accessed and other responsibilities related to it, what are the protocols which needs to be followed if you are accessing the data, all this information goes in the administrative part.

So basically, we see these parts, these five parts which are there in the metadata for

resource discovery. So, thick metadata is potentially providing indexing, access, annotation and classification for all data types including recording. So, indexing, access and annotation. And also classification.

So, everything is there in the thick metadata. It also carries the descriptive and analytical materials for the performances. So, if there is a song being performed, there will also be lots of information related to it, when the song is sung, who sings the song, who cannot sing the song. So, all these information is there the translation of it, annotation, everything is there in the thick metadata. Linguistic analysis which carries translation, transcription and other important descriptions are kept here. These might not be always useful for someone who wants to access data, may be someone wants to access data and only wants to know about the data.

Recordings provide evidence for analysis and make the descriptive and analytical process transparent and accountable. So, these evidences actually make the data transparent, so that one can know about it, and one also knows who have collected the data that makes it more accountable. So, what do we have? You can see again a list here, you can see speaker's full name. So, we have that language name, date of creation, use the primary equipment what was used while recording the data, date for the creation, place of the creation.

Now, all these information is there. Language, the place where it was collected, any restrictions were there and there can be also genre keywords dependent on the choice of the schema. So now, what does choice of the schema mean? There can be different types of schemas used in different projects and every schema has their own choice of words. As I said, different types of acronyms can be used, different types of fields can be used. So, depending on which schema one is following, it will it will change.

So, some of the popular schema are IMDI or OLAC schema. Label every metadata entry with the same level you use for the resource. So if you are following like OLAC schema, then you will follow it across the project in every language and list, every related item in the metadata. So, everything is listed. So you cannot say that I create metadata only for words and sentences, not for the narrations, not for folklores or oral literature, for everything actually it is being created.

So, this is a sample where you can see from IMDI. So, this is their way of creating metadata for language archive. They have mentioned title, date, place and description that information needs to be there, depositor's name and contact info, who is actually depositing the data, and contact details, project name, director, sponsor. So basically if I am working in a project, I can also upload my data for this archive with all these

information. Role, demographic data and their contact. So, who actually participated, who are the language experts, their role and demographic data, how many people are there and all those.

Resources like provenances, formats, relations, content like context, genre, narrative description, what is the content? And of course, references means any particular publications which we do find in that particular language. So, publications which are already available and then there is this OLAC data, where you can see that they have a different format. So, you can compare both these formats; they are little different, but then they all have basically the same information. So, there are contributor's or creator's name, title, date, description, resource info, relation to other objects, subject and then linguistic sub-field, type, that is genre.

So, all these things are there. So sometimes also, fields are arranged in different way or the field names are different in different catalogs or different metadata types, but basically we see that certain information are common across all these metadata. So now, this is a table where you see that when we are talking about archive, a particular format is used to store it because in archive, we are basically storing the data. And then when we present it, we use another format and when we are working on something, we use another format. So in archives generally, XML format is used for text, for photo TIFF and for audio WAV, while for presenting it, that means if you are searching somewhere and it gets in front of you, so if there is a description, it will come in PDF or HTML format. When you are searching for a particular description, you are searching for say, marriage songs in a particular language.

The format in which it is saved will not be visible to you, what will be visible to you is maybe in a PDF format or HTML format and the picture will also be visible to you in another format and same for the sound files. While working also, we can use another format. So, maybe we are writing in Word or Writter or any other Latex, anything else, but then it is stored in another format and it is made visible or presented in another format. Now, why I am talking about this because to make you understand that there are three levels. At one level, we work in the file and then at another level, we keep it preserved for a longer period of time; it is preserved in a folder.

And then there is another dimension to it, where we try to make it visible for those who want to look at it. If someone is searching for a particular thing, that particular data should be visible to that person. So, these three steps are involved. So, why to create and distribute metadata? Because it is very important for management of an archive. As I told that if the data is not systematically being kept, then how can one look into it, how can it be searched.

So, large archives almost always have internal metadata standards, they follow their own metadata standards which outsiders may not know, but outsiders can easily look at a particular data when he or she is looking for it, searching for it. It is also important for increasing awareness of your resources. So if the data is visible, then people will know about it, people will be aware about it. Using a standard metadata format actually helps because it will be easier to search for those data because they are following a standard format.

And also it increases the value of community-wide metadata resources. So, we know that there are 5 persons or 5 projects working on a particular language. If those projects are not creating their metadata, if those data are not visible, then no one will know whether those languages are actually being documented or not. So, because you know language documentation is not a one way research, it has also to give some outcomes to the community. So if some data is being collected, then outcomes can also be created or some interdisciplinary works can be done.

So, it is always good to make that process visible that data visible. So, the whole process of language documentation is changing with time and along with the creation of metadata, it is getting lots of importance. So, we see lots of people talking about actually creating metadata and how to organize it in a more systematic manner. Metadata is actually an age old practice or concept where we earlier had catalogs in the libraries and all. But in today's world, what we are doing is we are doing it in a digital format and we are trying to make it more systematic and also searchable because we have the option of search engine. It is important for systematic data preservation and for more visibility of the particular work.

Metadata are important for both the huge projects where lots of people are working in different languages, and it is also important if you are working alone in one language, because if you are keeping your data in a very haphazard manner anywhere in the cloud or in your PC, you will not be able to find it or it will take lots of time to search for anything if you want to go back. Because of course, language documentation takes lots of time; it is not done in one day. So, if you are investing 2-3 years working on it and suddenly, you want to go back to an older data, you might waste lots of time searching for it if it is not organized in a proper manner. That is why it is very important to create metadata also for the individuals who are working, maybe in one aspect of a particular language.

So, I hope you enjoyed today's class. Please go through these references, this will be very helpful. Thank you!